AFRL-RI-RS-TR-2013-096

UNDERSTANDING TONAL LANGUAGES

BINGHAMTON UNIVERSITY

APRIL 2013

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2013-096 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /

STANLEY J. WENNDT
Work Unit Manager

/ S /

WARREN H. DEBANY, JR.
Technical Advisor, Information
Exploitation & Operations Division
Information Directorate

REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* APRIL 2013	**2. REPORT TYPE** FINAL TECHNICAL REPORT	**3. DATES COVERED** *(From - To)* NOV 2010 – NOV 2012

4. TITLE AND SUBTITLE UNDERSTANDING TONAL LANGUAGES	**5a. CONTRACT NUMBER** FA8750-11-1-0061
	5b. GRANT NUMBER N/A
	5c. PROGRAM ELEMENT NUMBER 65502F
6. AUTHOR(S) Stephen A. Zahorian	**5d. PROJECT NUMBER** E2TL
	5e. TASK NUMBER BA
	5f. WORK UNIT NUMBER 01
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Department of Electrical and Computer Engineering Binghamton University Binghamton, NY 13902-6000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIGC 525 Brooks Road Rome NY 13441-4505	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL/RI
	11. SPONSOR/MONITOR'S REPORT NUMBER AFRL-RI-RS-TR-2013-096

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report gives a detailed summary of research work completed under Air Force Research Laboratory (AFRL) grant 56236, over the time period (November 17, 2010 – November 16, 2012). The main objective was to study various methods for Mandarin syllable recognition. Techniques were explored for both base syllable recognition and lexical tone recognition. The RASC863 database, obtained from the Chinese Linguistic Data Consortium was used for experimental work. Basel syllable phone recognition (60 phones) was done with a Hidden Markov Model recognizer. Best results obtained were approximately 69%. Human listeners were used to establish a baseline for lexical tone recognition. Tone recognition accuracy for humans ranges from about 55% to about 90%, depending on how much context is given to the listeners. The best tone classification accuracy with a neural network classifier is about 76%. The best tone recognition accuracy obtained with a Hidden Markov Model recognizer is about 71%. In addition to ASR experiments with Mandarin, basic research on improved pitch tracking, and refinement of spectral/temporal features (DCTCs/DCSCs) was done. It was determined that much longer time intervals are preferred for dynamic feature calculations than are typically used with MFCC features. Also the “best” segment intervals for Mandarin feature calculations are somewhat longer than for English.

15. SUBJECT TERMS
Open-source Videos, Mandarin Speech Recognition, Tonal Languages

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** **STANLEY J. WENNDT**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	59	**19b. TELEPONE NUMBER (*Include area code)*** **N/A**

Table of Contents

Section **Page**

List of Figures

List of Tables

1 Summary

This report gives a detailed summary of research work completed under Air Force Research Laboratory (AFRL) Award No. FA8750-11-1-0061, "Understanding Tonal Languages," over the time period (Nov 17, 2010 – Nov 16, 2012).

2 Project Introduction

Automatic speech recognition systems for tonal languages have been studied for many years. Most research targets Mandarin Chinese for investigation; due to its very large size and world-wide importance (845 million native and 1025 million total speakers [1].) There are two major barriers to progress. One is the lack of a large database that provides transcriptions finer than the syllable level. Another problem is that there are large differences among regional dialects of Chinese so that an enormous database with a very a large number of speakers would be required to develop a general Mandarin ASR system, without degradation due to accent issues. To date, nearly all high accuracy systems have been developed using either isolated words (monosyllables) with read speech from broadcast news announcers or synchronized computer voices.

In this work, we have been extensively investigating signal processing and recognition strategies for Mandarin syllable and tone recognition. Unlike English, Mandarin is composed of approximately 416 distinct base syllables, which, in total, represent approximately 1345 tonal syllables (Wang et al., 1995) [2], using five types of pitch contours (tones) for further differentiation of the base syllables. Mandarin Automatic Speech Recognition (ASR) has not yet been as intensively investigated as English language ASR. However, since at least 1995, considerable effort has been devoted to base syllable recognition and tone recognition. Robust ASR for Mandarin is potentially of even more significance than English ASR, because of the great difficulty of inputting Mandarin characters to a computer via keyboard, the very large number of people who speak Mandarin as their first language and the great interest in automatic real time spoken language translators between Mandarin and other languages such as English.

The vast majority of reported work for Mandarin base syllable recognition utilizes Mel cepstral coefficients as the primary acoustic features for a Hidden Markov Model (HMM) and/or neural network recognizer. Much of the work on tone recognition has focused on developing better pitch trackers, due to the importance of robust pitch tracking for tone recognition, and the perceived deficiencies in existing trackers. In the work described here, the recently introduced type of spectral/temporal DCTC/DCS features was used for the baseline acoustic features (Zahorian et al., 1997) [3]. Although similar generally to Mel cepstra and delta cepstra, the DCTC/DCS features have considerably more flexibility for trading off between spectral and temporal resolution. We have also shown these features to be superior to the more standard Mel cepstra features in controlled comparison tests for phonetic recognition in English. For the case of Mandarin monosyllables, the DCTC/DCS features can be further optimized to take into account such characteristics as the syllable “FINAL” having more influence on the syllable “INITIAL” than vice versa. For the pitch tracking needed as the first step for tone recognition, the YAAPT tracker was used (Zahorian and Hu, 2008) [4]. This tracker has been shown to be very accurate, performs very well in the presence of noise, and can also easily be adjusted to either provide a continuous pitch track, smoothly spanning unvoiced speech, or reliably make voiced/unvoiced decisions. Both modes of operation were evaluated in the research for syllable and tone recognition. However, generally the continuous pitch track mode (speech is assumed to be all voiced) is preferable.

Beginning with the front end DCTC/DCS acoustic features, and pitch contours, a combination neural network/HMM recognizer was used to evaluate a variety of methods for computing and combining these features. For example, a multiple pass procedure can be used whereby the approximate locations of each syllable and the parts of each syllable are estimated on the first pass, and then feature calculations are redone on a second pass, synchronized to the estimated “INITIAL” and “FINAL” locations and adjusted to take into account anticipated co-articulation effects. Experiments reported here were conducted with the RASC863 project Mandarin database, for which “ground truth” syllable and tone labels are available. There are also published results for base syllable, tonal syllable, and tone error rates, which can be compared for judging the effectiveness, and areas needing improvement, in the work done over the funding period.

Although tone recognition has been investigated for many years, relatively high recognition accuracy has only been obtained for isolated words and read speech (Huang and Seide, 2000)[5]. Various pattern recognition methods were applied to tone recognition, including HMMs, Gaussian Mixture Models (GMMs), and Decision-tree Classification. Approaches in the tonal language recognition fall into two major categories: embedded tone modeling and explicit tone modeling (Liu et al., 2007) [6]. In embedded tone modeling, pitch related features are added as extra dimensions in the short time speech feature vector. Tone recognition has been done as an integral part of the existing system (Chang et al., 2000) [7]. For example, Zhou et al. (2004) [8] combined MFCC and pitch based features into one feature vector for tone modeling. In addition, the pitch duration and long span pitch information were integrated into the feature vector. In contrast, in explicit tone modeling, tones are independently modeled and recognized in parallel to the recognition of base syllables. In the work of Wang et al. (1995) [2], a set of sub-syllabic models for base syllable recognition and another set of context-dependent models for tone recognition were employed, and then the results were synchronized by a concatenated syllable matching algorithm.

Furthermore, Wang and Seneff (2000) [9] have investigated the interaction between intonation and tones in Mandarin and found the tone classification errors were reduced when the effects of tone coarticulation and intonation were normalized. Meanwhile, spontaneous speech recognition is also a big challenge in Mandarin recognition because spontaneous speech usually contains mispronunciations, emotional status, and a fast speaking rate (Liu et al., 2007) [6].

3 Methods, Assumptions and Procedures

3.1 Overview of MANDARIN CHINESE

3.1.1 Language Elements and Structure

3.1.1.1 LANGUAGE STRUCTURE

The basic unit of Mandarin Chinese is a character. Typically, a meaningful Chinese sentence is built from words - even though the sentence can also seen as a string of characters. Each character is pronounced as a syllable. Each character, however, can correspond to more than one syllable (more than one way to be pronounced). The relation among syllables, phones, and tones is described in later sections.

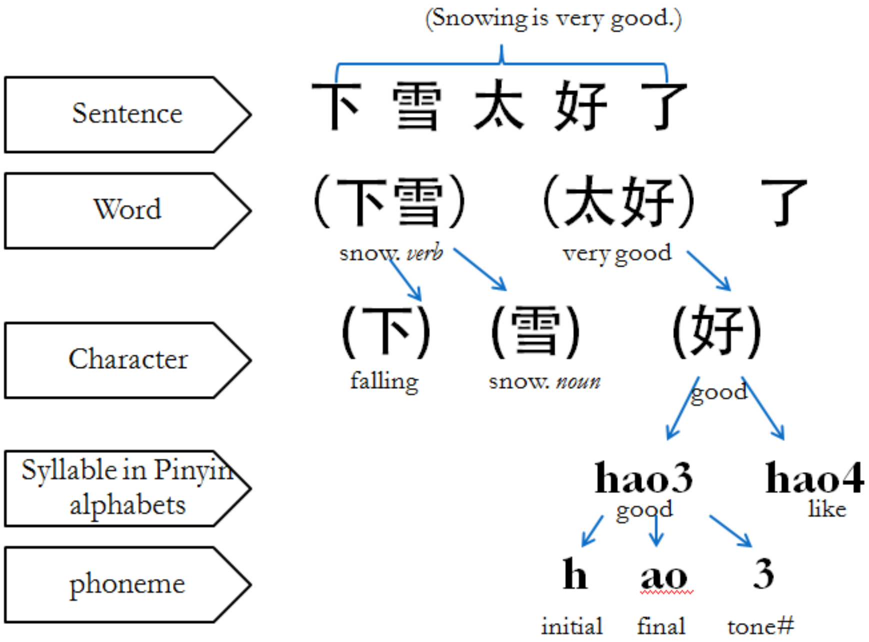


Figure 1: Language structure of Mandarin Chinese

3.1.1.2 SYLLABLES AND PINYIN

As a tonal language, Mandarin uses a mono-syllable associated with a lexical tone as the basic unit. If the tone is neglected, the unit is referred to as a base syllable.

Conventionally, each Mandarin base syllable can be decomposed into an "INITIAL/FINAL" format very similar to the consonant/vowel relations in English. "INITIAL" is the initial consonant part of a syllable and "FINAL" is the vowel part but including optional medial or nasal ending. There are a total of 22 "INITIALs" and 41 "FINALs" in Mandarin (Wang et al., 1995) [2].

3.1.1.3 LEXICAL TONES

For each syllable, there is a lexical tone indicated by syllable-level pitch or fundamental frequency (F0) contour patterns. Because pitch is only defined for the voiced regions, it is generally assumed that the tone is associated only with "FINALs." There are five tones in Mandarin: high-level (tone 1), high-rising (tone 2), low-dipping (tone 3), high-falling (tone 4) and neutral tone (tone 5). In actual speech, the neutral tone is rare, and is sometimes judged as "no tone." Figure 2 shows the wave, energy and pitch of basic syllable "ma" with 4 different tones.

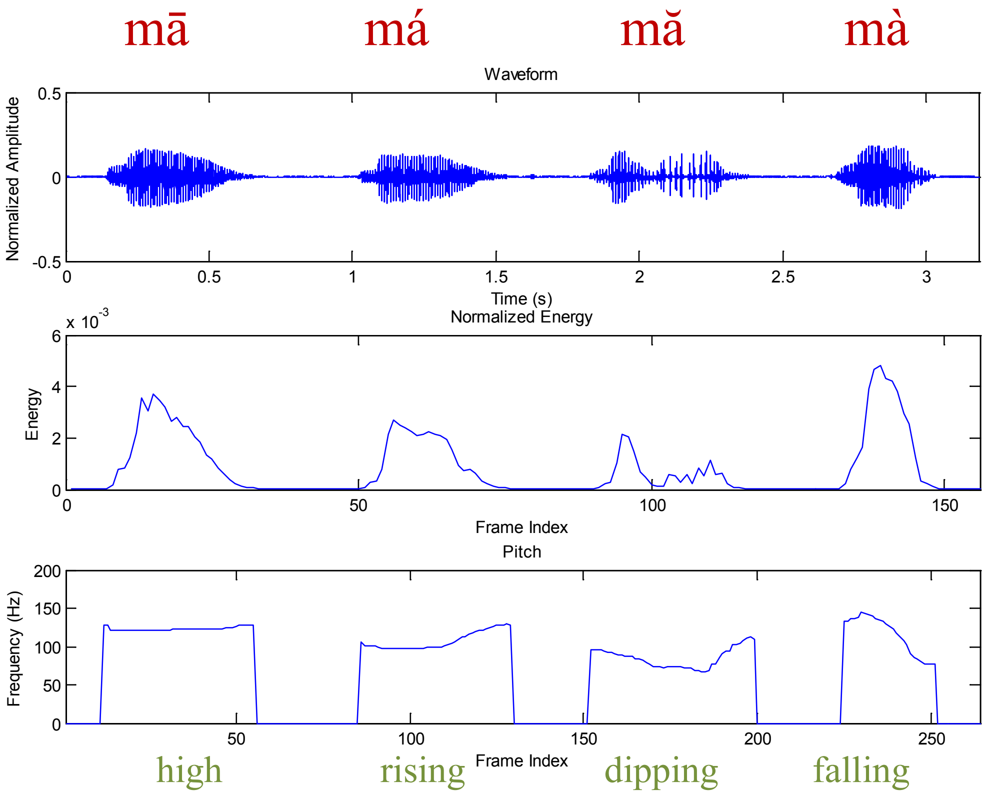
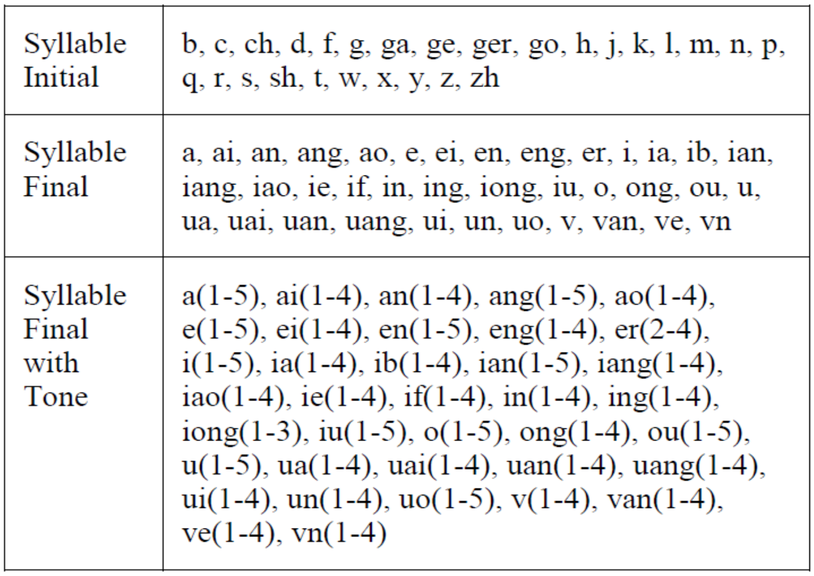


Figure 2: Illustration of Mandarin tones

3.1.1.4 CONCLUSION

The total number of phonologically allowed different syllables in Mandarin is 1345. However, when the tone information is disregarded, there are only 416 base syllables (Wang et al., 1995) [2]. Therefore, accurate tone recognition plays an important role in automatic Mandarin speech recognition and a large amount of research has been devoted to the pursuit of an accurate pitch tracking algorithm (Lin and Lee, 2003 [10]; Huang and Seide, 2000 [5], Zahorian and Hu, 2008 [4]). Table 1 lists typical Syllable INITIALs and FINALs and all legal combinations of Syllable FINALS and tones.

Table 1: Syllable INITIALs and FINALs and all legal combinations



Syllable Initial	b, c, ch, d, f, g, ga, ge, ger, go, h, j, k, l, m, n, p, q, r, s, sh, t, w, x, y, z, zh
Syllable Final	a, ai, an, ang, ao, e, ei, en, eng, er, i, ia, ib, ian, iang, iao, ie, if, in, ing, iong, iu, o, ong, ou, u, ua, uai, uan, uang, ui, un, uo, v, van, ve, vn
Syllable Final with Tone	a(1-5), ai(1-4), an(1-4), ang(1-5), ao(1-4), e(1-5), ei(1-4), en(1-5), eng(1-4), er(2-4), i(1-5), ia(1-4), ib(1-4), ian(1-5), iang(1-4), iao(1-4), ie(1-4), if(1-4), in(1-4), ing(1-4), iong(1-3), iu(1-5), o(1-5), ong(1-4), ou(1-5), u(1-5), ua(1-4), uai(1-4), uan(1-4), uang(1-4), ui(1-4), un(1-4), uo(1-5), v(1-4), van(1-4), ve(1-4), vn(1-4)

3.1.2 Language Modeling

As tones are a very important characteristic for a tonal language and a primary difference with non-tonal languages such as English, there are two major approaches for building an ASR system for Mandarin Chinese speech: recognition with implicit tone modeling (ITM), or recognition with explicit tone modeling (ETM) [6].

For embedded tone modeling, pitch related features are added as extra dimensions to the short time spectral speech feature vector. Recognition is done as an integral part of the existing system [11]. For example, Zhou and et al. [12] combined MFCCs (Mel-Frequency Cepstral Coefficients) and pitch based features into one feature vector for embedded tone modeling.

On the other hand, in explicit tone modeling, tones are independently modeled and recognized using prosody features [13] in parallel with the recognition of base syllables, which ignore tone numbers.

For ASR at the phonetic level, the required number of HMMs for English, Mandarin modeling using ETM, and Mandarin modeling using ITM are listed in table 2. For the case of Mandarin, the number of HMMs for various possible “versions” of both ETM and ITM are listed. Note, that in contrast to English, where only 48 HMMs are needed (and in practice, usually only 39), a very large number of HMMs are needed for Mandarin (up to 1345 for one approach). This implies, at least potentially, much more speech data is needed for HMM modeling of Mandarin then for English. (However, it should also be noted that some of the most accurate English ASR systems use triphones as fundamental HMM building blocks, and there are typically on the order of 1000 triphone HMM models.)

Table 2: Language modeling for English and Chinese

ENGLISH	MANDARIN CHINESE			
	Explicit tone modeling (ETM)		Implicit tone modeling (ITM)	
48 HMMs	421 HMMs	68 HMMs	1345 HMMs	187HMMs
48 HMMs for 48 phones	416 HMMs for 416 distinct syllables. Plus 5 for recognizing tones	For 22 INITIALs and 41 FINALs plus 5 for recognizing tones + simple grammar	One HMM per syllable for all 1345 tonal syllables	22 HMMs for 22 INITIALS, and 165 HMMs for all legal tonal FINALs + simple grammar

3.2 DATABASE: RASC863 Shanghai Region

3.2.1 Introduction

The main database that was used for experimental results reported here is RASC863, which is a multi-speaker Mandarin speech database developed by the Chinese Linguistic Data Consortium (CLDC) [14]. Summarizing briefly, the entire RASC863 annotated database is a speech corpus with 4 regional accents, and is a collection of read and spontaneous speech from 800 speakers equally spread among 4 major regions (Shanghai, Chongqing, Fujian, and Guangzhou) in China (200 speakers for each region.) The transcriptions of the speech for 20 speakers from each region are annotated at the phonetic level. Totally the database contains about 34 hours of speech, or 8.5 hours per region.

For the work summarized in this report, the speech from the Shanghai region was chosen based on the consideration that the accent for speech from this region is most "neutral" out of the 4 regions. The data from each speaker includes one 3 to 5 minute long spontaneous speech passage, 15 spontaneous short sentences, 20 short read sentences and roughly 110 phonetically balanced sentences (the "S-group"). The cumulative training time of the phonetically balanced part is similar to that of the TIMIT database [15]. A comparison between RASC and TIMIT is listed in Table 3. The RASC863 database also provides 4 different types and three different resolution levels of labels for each sentence. Each sentence is transcribed as the entire sentence using Chinese characters; again at the sentence level but using PINYIN; with syllable boundaries marked for each syllable (PINYIN); and with phoneme boundaries marked (PINYIN).

Due to the fact that the database chosen for this research only contains speech from 20 speakers (phonetically transcribed) from Shanghai and, from casual listening, there appears to be considerable accent variations among the speakers (despite the earlier comment about “relatively” accent free), there are not enough speakers to perform a meaningful completely speaker independent evaluation with separate training and test speakers. For example, for RASC863 14 speakers could be used for training, with the remaining 6 speakers for test, whereas for the TIMIT database, 3696 sentences (~191 minutes) from 462 speakers can be used for training, and 1344 sentences (~70 minutes) from 168 speakers for test. Thus, despite the overall large duration of speech used in the RASC863 data, the small number of speakers are likely not large enough to generalize well to other speakers. Therefore, for experiments reported in this paper, 70% of the phonetically balanced sentences (1548 sentences, ~224 minutes) from every speaker were used as training data, and the remaining 30% (661 sentences, ~96 minutes) were used for testing data.

Table 3: Comparison of RASC863 (Mandarin) with TIMIT (English)

	TIMIT	**RASC863**
Speakers	630	20
Speech type	Read	Answers to questions, read
Time duration	400 minutes	510 minutes 178 sentences per speaker
Label types	*.TXT - sentence level *.WRD - word level *.PHN - phoneme level	*.TextGrid file includes sentences level in both char/syllable; include syllable level; include phoneme level
Accent	Divided into 8 dialect regions	All speakers from Shanghai
Non-speech notations	No	Yes (e.g. [BR] – breathing) is labeled, too.
Sampling Rate	16 kHz	16kHz

3.2.2 Database Corrections and Modifications

3.2.2.1 CORRECTIONS FOR GENERAL ISSUES

In all our previous acoustic-phonetic ASR work using the HTK, the TIMIT database was used. Therefore, in order to avoid HTK malfunctions when using the same conventions for setup files, the labeling files for RASC863 had to be formatted the same as were the TIMIT labeling files. Several changes in the labeling files for RASC863 were required. For example for the RASC863 database, the start and end time for each syllable had to be converted to sample indices; the original tones labels (e. g. numerical digits 0, 1, 2, 3, 4 for the 5 lexicon tones) had to be either removed (for experiments involving base syllables) or converted to an alphabetic notation. Also, the phonetic transcription part of RASC863 includes human introduced noise or mispronunciations. Some notations used for transcribing these special events are listed in Table 4.

Table 4: Special notations used in RASC863

Sound change		**Mispronunciation**		**Human noise**	
Type	Symbol	Type	Symbol	Type	Symbol
Nasalized	~	Dialectic	#	Sniffle	[SN]
Centralized	_”	Pronunciation	*	Deglutition	[DE]
Voiced Consonant	_v			Inspiration	[IP]
Breathy	_h			Yawn	[YA]

It would have been difficult, and somewhat of a distraction to the main goals of this work, to explicitly model these noises and other “problems.” So their corresponding transcriptions were modified with rules as defined in the following table:

Table 5: Modification for special notations

Special Sound	**Symbol**	**Modification**
Nasalized	~	Ignored if with a phone/”Sil” if appears alone
Centralized	_”	Ignored if with a phone/”Sil” if appears alone
Voiced Consonant	_v	Ignored if with a phone/”Sil” if appears alone
Breathy	_h	Ignored if with a phone/”Sil” if appears alone
Dialectic	#	Ignored if with a phone/”Sil” if appears alone
Pronunciation	*	Ignored if with a phone/”Sil” if appears alone
Sniffle	[SN]	Change to “Sil”
Deglutition	[DE]	Change to “Sil”
Inspiration	[IP]	Change to “Sil”
Yawn	[YA]	Change to “Sil”

3.2.2.2 MODIFICATION FOR PHONETIC TRANSCRIPTION

Due to the fact this research used only the Shanghai region data of the entire RASC863 database, there are special phones that are only used in the “Wu” language (e.g. accented Mandarin type in Shanghai region) that can be found in the transcriptions. However, as judged by human listeners, the accented phones sound very similar to their equivalent phones in “standard” Mandarin. They also only appear several times in the entire database. Thus the transcriptions of these special phones were converted to their phonetic equivalence in “standard” Mandarin Chinese.

Table 6: Modification for special phones of “Wu”

Phone Only Used By “Wu”	**Equivalent Phones in Mandarin**
Oong	Ong
Voong	Vn
Eu	Ou

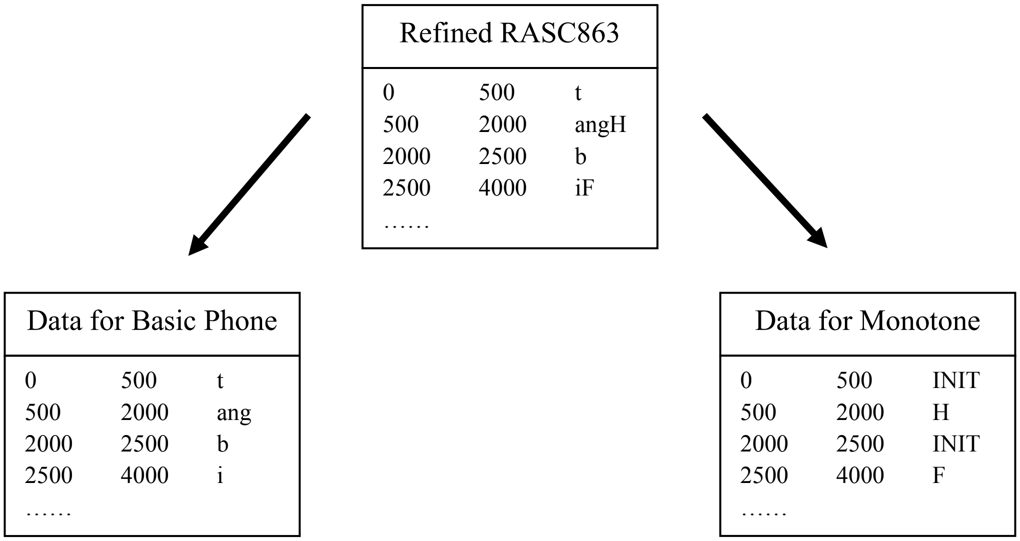
Also there were typo errors found in the phonetic transcription from this database. They had to be corrected before running with HTK. Otherwise, the HTK failed to generate models for several cases due to insufficient data. That is the HTK attempted to generate a model for each type of typo—meaningless models even if they could have been generated. Typical error types and how they were corrected are summarized in Table 7 below:

Table 7: Error patterns in RASC863

Error Type	Example	Correction
Typo	ven, veng	Correct with right phone label "uen"
Combined	zhi	mis-combined the labels "zh" and "i" from two lines into "zhi"
Empty Transcription	No label	Usually appears with "Combined" error type
Forget Tone Number "0"	e	Put a "0" at the end

Note that to meaningfully use the remainder of the RASC863 database, similar kinds of errors should be removed. Unfortunately, this error removal process is tedious and time consuming, as it is very difficult to completely automate this "clean-up" process.

Data used for research on basic phone recognition, and for phone recognition and tone recognition, were all derived from the refined version of the original RASC863 from section 4.2., except with a different transcription format. This process is illustrated in Figure 3.



Refined RASC863

```
0       500     t
500     2000    angH
2000    2500    b
2500    4000    iF
……
```

Data for Basic Phone

```
0       500     t
500     2000    ang
2000    2500    b
2500    4000    i
……
```

Data for Monotone

```
0       500     INIT
500     2000    H
2000    2500    INIT
2500    4000    F
……
```

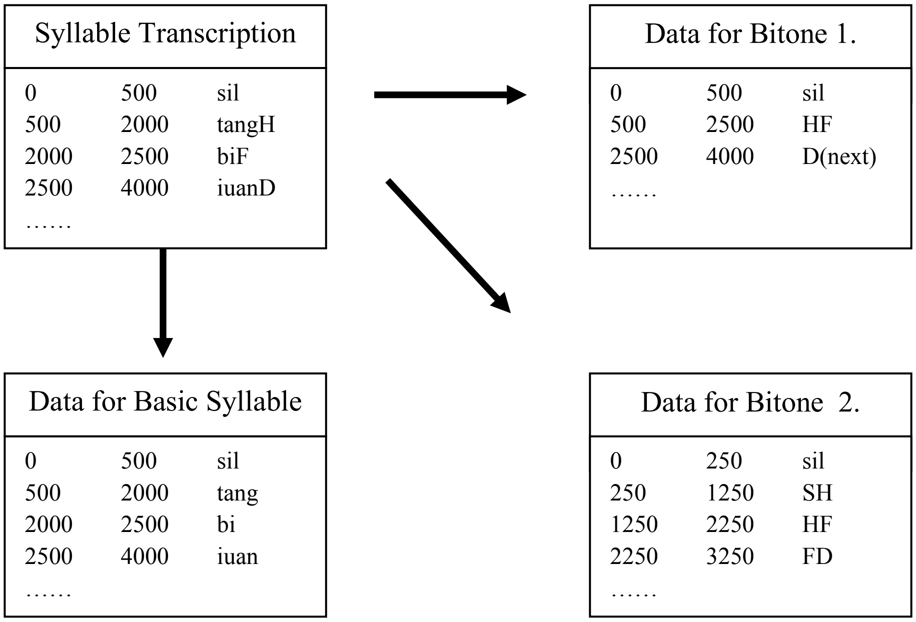
Figure 3: Data modification for RASC863 (1)

3.2.2.3 MODIFICATIONS FOR SYLLABLE TRANSCRIPTIONS

The transcription at syllable label is more precise since only very few typos were found. Thus, the original data could be used for recognition at the syllable level after suppressing tone numbers. After data modification and preparation, the data can be used for explicit tone recognition of tone pairs (Bitones).

The idea of the modification is to label the time when a speaker starts transitioning from one tone to the next one without "breaking" each transcription Two methods have been attempted for this purpose. In method 1, two consecutive tones are grouped together (H in tangH and F in biF) to create the longer Bitone (HF). However, monotone segments followed by silence are retained as monotones in in the modified labeling scheme. For method 2, each Bitone begins at the half way point of the "first" tone (even if silence) and ends at the halfway point of the "second" tone (even if silence). Thus there are no monotone segment labels for method 2. This process is also illustrated in the figure below.

Automatic recognition of these Bitones is described later in this report (section 6.3).



Syllable Transcription

0	500	sil
500	2000	tangH
2000	2500	biF
2500	4000	iuanD
……		

Data for Bitone 1.

0	500	sil
500	2500	HF
2500	4000	D(next)
……		

Data for Basic Syllable

0	500	sil
500	2000	tang
2000	2500	bi
2500	4000	iuan
……		

Data for Bitone 2.

0	250	sil
250	1250	SH
1250	2250	HF
2250	3250	FD
……		

Figure 4: Data modification for RASC863 (2)

3.3 ASR for Continuous Mandarin at Phonetic Level

3.3.1 Introduction

The experimental setup for performing base phone (that is, ignoring the tone number) recognition is very similar to that for phonetically recognizing 48 phones in English.

Table 8 lists all Mandarin phones used for these experiments. A major way to separate Chinese phones is to categorize them into “Syllable Initials” and “Syllable Finals.” Notice that elements listed for “Syllable Initials” are mostly consonants while the “Syllable Finals” mostly include vowels and vowels ending in a nasal sound. Spoken Mandarin Chinese is a tonal language and most of its syllables are meaningful only if tones are also known. Thus, the RASC863 transcription also includes tone numbers after nearly every vowel. Therefore, the very beginning step for performing base phone recognition with RASC863 was to remove these tone numbers in the annotation files. With tones ignored, there are 58 basic Mandarin Chinese phones, each of which was modeled by a 3-state HMM. Additionally, one HMM was created for silence and one was created for all sounds labeled as “human noise and mispronunciation.” Thus, a total of 60 HMMs were used for these experiments, versus 39 HMMs typically used for English.

Table 8: Required HMMs for basic phone recognition

Syllable Initials	Syllable Finals	Silence
z, zh, j, m, q, p, sh, d, x, l, g, b, ch, h, k, t, n, f, c, r, s	sil, ong, ian, u, ie, an, uan, iou, ing, ang, i, iii, ai, van, iang, ou, iao, ao, eng, ua, uo, uei, a, en, e, ei, v, iong, uen, ia, uai, ii, ve, uang, o, in, er, vn, eer, ueng	Sil

3.3.2 Spectral/Temporal Features

Although most ASR research uses mel frequency cepstral coefficients and delta terms as acoustic features, we have found that a somewhat different (although not too different) feature set called DCTCs and DCSCs have performance advantages over the MFCC features [34]. Essentially the DCTCs are computed similarly to MFCC terms, except no filter bank step is used; the mel frequency scale is realized with modifications to the cosine basis vectors used. The DCSC terms are conceptually similar to MFCC delta and delta terms, except much longer time intervals and more terms are used as compared with the MFCC delta and delta-delta terms. Although the total number of DCTC/DCSC terms can easily be varied, and often is depending the details of the ASR task. For this task, we use 13 DCTC terms, each encoded with 5-6 DCS terms (65-78 total features, and sliding blocks of length approximately 200 ms. In contrast the standard for MFCCs is a feature vector of length 39 (12 MFCCs plus energy, 13 delta terms, and 13 delta-delta terms), using effective blocks lengths of about 45 ms.

Our research on DCTC/DCSC features over the past few past few years has also shown that the temporal aspect of the acoustic features are quite important relative to spectral resolution [16]. Best results in both noisy and clean conditions are typically obtained with initial spectral analysis window lengths of about 8-10 ms, a frame spacing of about 2 ms, and a block length of about 200 ms for computing DCSC terms (or on the order of 100 frames).

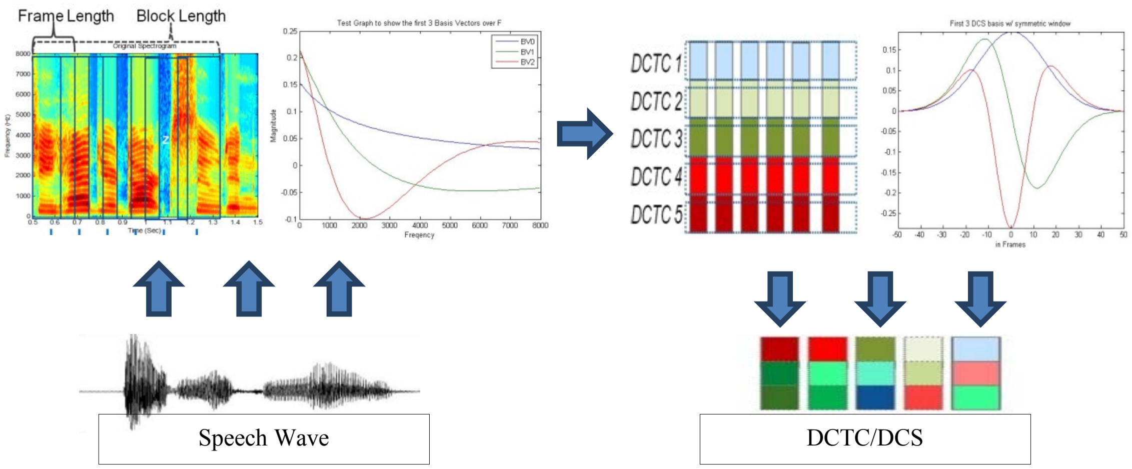


Figure 5: DCTC/DCSC feature extraction

A major objective of the present work was to explore time/frequency resolution for Mandarin, using the DCTC/DCSC features which easily allow tradeoffs in time and frequency resolution. Note that Chinese speakers usually speak 160-180 characters per minute which equates to 160-180 syllables per minute since all Mandarin Chinese characters are monosyllabic. This result in a much lower syllable rate compared to the average American adult who reads prose text at 250 to 300 words per minute - with an even larger number of actual syllables. The working hypothesis was that these time/frequency effects are quite different for Mandarin versus English.

3.3.3 Base Syllable Recognizer

Like much other research in language processing, HMM acoustic models were used for the ASR system. HTK (Version 3.4) was the actual software system used. This toolkit allows complete flexibility in terms of the number of mixtures, number of states, types of transitions, and provides for language modeling.

3.3.4 Experimental Evaluation

The objective of these experiments was to determine the accuracy possible using DCTC/DCSC features for recognizing Mandarin phones without considering tones (i.e, tone number). Again, the standard defined in Table 8 will be used through all experiments in this section. The accuracy achieved using control MFCC-39 features (that is, 12 MFCCs plus energy with delta and acceleration terms, or 39 total terms were obtained every 10 ms at a frame length of 25 ms, with pre-emphasis coefficient of 0.97) was 68.8%, which is roughly 1.5% higher than obtained with 39 DCTC/DCSC features. When increased to 48 features, MFCC (68.8%) and DCTC/DCSC gave very similar results. The best accuracy (70.4%) was obtained using 65 DCTC/DCSC features computed using a block length of 340ms for computing DCSC trajectory features.

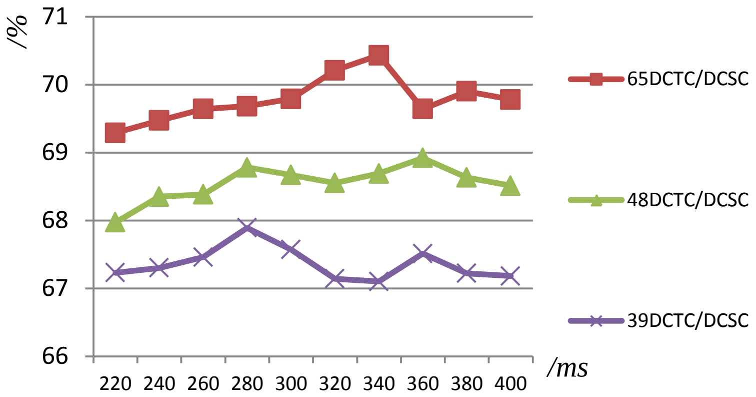


Figure 6: Basic phone recognition using DCTC/DCSC features and HMM phone models.

Another experiment compared the accuracy of English phone recognition (TIMIT) and Mandarin Chinese basic phone recognition (RASC863), using DCTC/DCSC features, as a function of trajectory length. Notice Mandarin speech has 60 phone models that need to be trained with a similar amount data as that used to train 39 English phone models, so using the same feature configuration (48 DCTC/DCSC), a somewhat lower accuracy might be expected for Mandarin than for English (Figure 7).

This experiment also shows that the "best" DCSC block for Mandarin Chinese is longer than the "best block length for English. It is clear to see that the maximum recognition accuracy (72.5%) for English is obtained at 300 ms. This block length is about 60 ms shorter than where the peak value of 68.9% for Mandarin Chinese recognition is obtained (360 ms). This phenomenon supports the hypothesis that the length of time for computing spectral trajectory features for Mandarin should be longer than the length of time for computing spectral trajectory features for English.

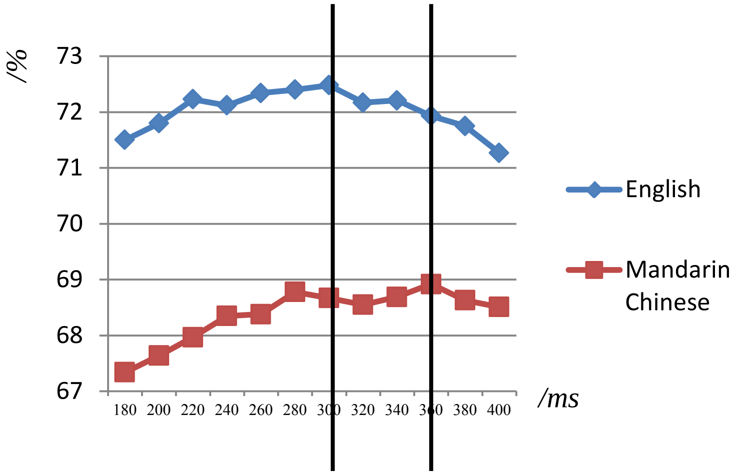


Figure 7: Phone recognition using TIMIT vs. basic phone recognition using RASC863

3.3.5 Conclusions

In the method presented in this section, acoustic features - DCTC/DCSC - are obtained by extracting the temporal trajectories of integrated frequency domain features and tested with the RASC863 database through ASR experiments. The results not only demonstrate that DCTC/DCSC features are effective for base phone recognition but also suggest that the temporal trajectory length for processing Mandarin should be longer than for English.

3.4 Lexical Tone Recognition in Continuous MANDARIN

3.4.1 Introduction

Most literature on machine recognition of tones is based on syllables spoken in isolation [17] or high quality speech such as TV broadcast news [18]. This is likely due to the fact that recognizing tones from syllables extracted from conversational speech is difficult even for humans: some linguistic research suggests that human listeners require long duration acoustic cues in order to recognize tones correctly [19]. The perception of tones also varies depending on the listener's native experience with the tonal system of his/her own language [20].

The following section of this research report explores the recognition ability of humans for lexical tones in lightly-accented continuous Mandarin Chinese for different conditions. Then two automatic methods for monotone recognition are discussed and compared with both human's accuracy and other tone modeling techniques [21][22].

The Shanghai region data from RASC863 (Regional Accented Speech Corpus) [14] was utilized for all experiments reported in this paper, since it provides phonetically labeled transcriptions and the accent from speakers in this region is considered the "lightest" of all 4 regions included in the RASC863 database.

3.4.2 Tone Recognition by Humans

3.4.2.1 INTRODUCTION

The main object for this experiment was to investigate the capability of native listeners to recognize tones, for each of four cases, with varying amounts of context (i.e., length). For each case, listeners were given approximately 800 speech segments to listen to and were asked to identify tones, based on a single playback of each speech sample. The segments (approximately 3200 in total) were directly extracted from the continuous speech data portion of the RASC863 database. The type of speech segment for each case is listed in Table 9.

3.4.2.2 EXPERIMENTAL PROTOCOL AND TEST SOFTWARE

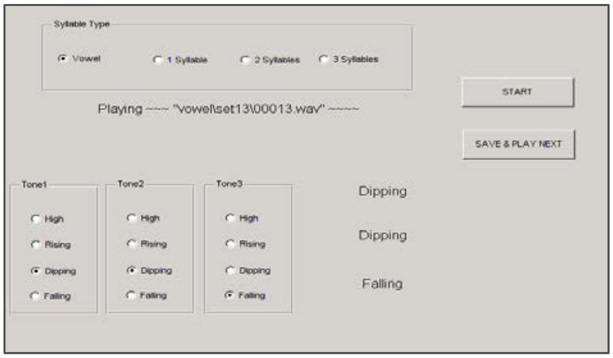
3 male and 3 female college students whose native language is Mandarin Chinese were selected as the experimental subjects. All participants were trained and tested by the grad student supervising this phase of the work (James Wu) and shown a correct understanding of lexical tones in Mandarin (High, Rising, Dipping and Falling) and their corresponding pitch characteristics.

The syllable segments for Cases 3 and 4 are consecutive syllables extracted from random positions in a sentence. In some cases the syllable strings were “words,” but not in all cases, as discussed later.

Table 9: Cases for human listening test

Case 1	**Case 2**	**Case 3**	**Case 4**
One syllable final: (vowel part only)	One complete syllable (consonant and vowel)	A string that contains two syllables	A string that contains three syllables

Listeners used a PC for playing sound tokens and recording their answers with interaction with the computer via a Graphic User Interface (GUI) as shown in Figure 8. Listeners simply marked tone type (High, Rising, Dipping and Falling) according to what they just heard. Listeners were comfortably seated and used high quality headphones.



Syllable Type

● Vowel ○ 1 Syllable ○ 2 Syllables ○ 3 Syllables

Playing ~~~ "vowel\set13\00013.wav" ~~~~

Tone1: ○ High ○ Rising ● Dipping ○ Falling

Tone2: ○ High ○ Rising ● Dipping ○ Falling

Tone3: ○ High ○ Rising ○ Dipping ● Falling

Dipping

Dipping

Falling

Figure 8: GUI for listening experiment

Listeners were only allowed to listen once only to each token, but could decide when to begin the playback of each token. Typically listeners required about 90 minutes for each of the four cases.

3.4.2.3 ANALYSIS OF RESULTS

The overall tone recognition results from 6 listeners are in table 10 (labeled by 'Male1' to 'Female3') with respect to 4 cases (labeled by 'Vowel' to '3 slbs'), and their average and standard deviation are shown in the following table.

Table 10: Tone recognition accuracy (%) for human listeners

	M1	M2	M3	F1	F2	F3	Avg.	S. Dev
Vowel	59.4	51.1	70.0	67.3	59.9	63.6	61.9	6.7
1 slb	58.7	53.8	64.7	53.6	57.4	59.8	58.0	4.2
2 slbs	75.7	74.4	81.3	80.8	80.3	76.8	78.2	3.0
3 slbs	80.7	67.8	86.1	83.1	80.8	78.8	79.5	6.2
Average	68.6	61.8	75.5	71.2	69.6	69.8	69.4	4.5

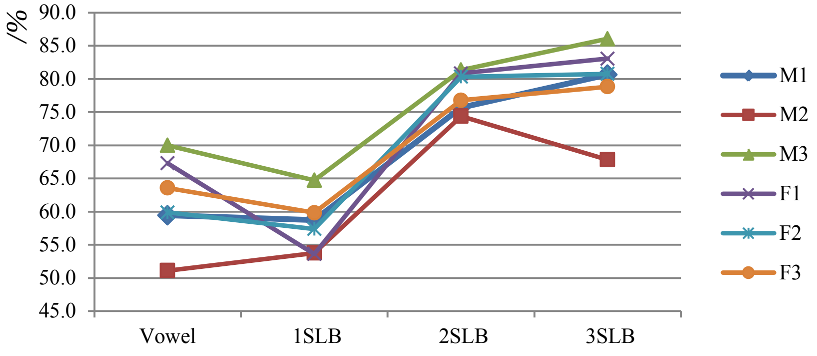


Figure 9: Tone recognition accuracy (%) for human listeners

The average recognition rate for case 3 and case 4 (e.g. string with multiple syllables) is noticeable higher (78.2% and 79.5%) than for the single syllable cases 1 and 2 (61.9% and 58.0%)

For most languages, the tonal effect only occurs in voiced speech. Thus most research on automatic Mandarin recognition, either implicitly or explicitly, modeled tones based on acoustic features that are extracted from vowels only.

The confusion matrix (Table 11) is generated from the recognition result for case "Vowel." All values in the table are per cents. The High tone appears most likely to be recognized correctly while Dipping and Rising tones appears to be more difficult to recognize correctly. This is consistent with [19], that finds that recognizing High tone and Falling tone requires shorter cues than for the other two tones (Dipping and Rising).

Table 11: Tone confusion matrix for human listeners based on listening to vowels only

	High	**Rising**	**Dipping**	**Falling**
High	70.4	10.9	5.4	13.3
Rising	21.7	59.6	8.6	10.2
Dipping	16.9	25.9	44.2	13.0
Falling	18.4	7.2	11.5	63.0

The next result is for case "2 slbs" only. Since each segment was randomly selected from a sentence, syllable pairs could form a lexical word, or might not form a word. (Most Chinese words are comprised of 2 characters (syllables)). As expected, when recognizing a string from 2 syllables that form a word, the linguistic information available to listeners helps them recognize tones much more accurately than for the case when they are recognizing tones from two syllables with no linguistic meaning.

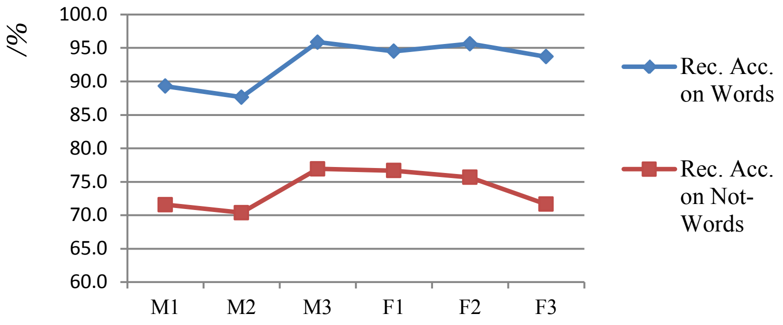


Figure 10: Recognition rate of tones for words versus not-words in two syllable cases.

The next result explores the possibility that there might be gender dependent effects for tone recognition. Results were rescored, sorting both listeners and speakers by gender. These results are shown in table 12. The basic conclusion is that there are no significant gender dependent effects for tone recognition. However the female listeners did score a little higher on tone recognition than did the male speakers (~5%). Tones from female speakers recognized slightly more accurately than tones from male speaker by both male and female listeners (1.9%). These observations may not be statistically meaningful, given the small pool of speakers (20) and listeners (3 female and 3 male).

Table 12: Tone recognition rate by gender of speakers and listeners

Spks/Lsns	**M**	**F**	**A**
M	75.1	78.5	76.8
F	75.8	81.6	78.7
A	75.4	80.0	78.2

The results depicted in Table 13, for case “3 syllables,” indicates that recognizing the tone of the middle syllable appears to be easier than the tone of the beginning or ending syllable for human listeners. This may be because the co-articulation cues for both the beginning and end part of the middle syllable are present (from first and third syllables) whereas the first syllable has only ending co-articulation cues (from the following middle syllable) and the last syllable has only beginning co-articulation cues (from the preceding middle syllable).

Table 13: Tone recognition rate as a function of syllable position in 3-syllable strings

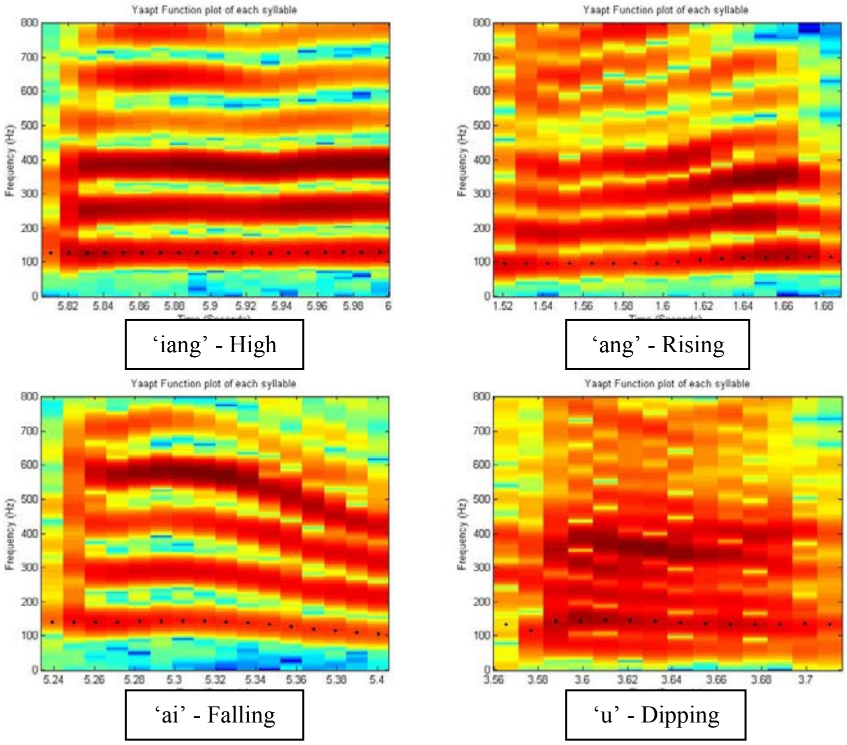
	M1	M2	M3	F1	F2	F3
1st slb	76.3	72.9	80.4	79.5	76.2	73.5
2nd slb	78.4	80.4	90.9	88.3	86.2	83.5
3rd slb	87.3	69.9	86.8	81.3	79.8	79.5

3.4.3 Tone Classification Using Neural Networks

3.4.3.1 PITCH FEATURE

The pitch contour is widely accepted as one of most effective features for tone tracking [17][20][21][22][14]. The research reported in this paper made use of the fundamental frequency tracking algorithm called YAAPT [4] for pitch. Pitch contours were represented with Discrete Cosine Series Coefficients terms (4-5) in the same manner as DCTC trajectories are encoded [16]. The black dots placed on the spectrograms shown in Figure 11 are pitch tracks computed by YAAPT.

The four panels in Figure 11 illustrate the capability of this method for pitch tracking. The black dots laying on the spectrograms for each tonal syllable is the pitch track.



'iang' - High

'ang' - Rising

'ai' - Falling

'u' - Dipping

Figure 11: Low frequency part of spectrogram with pitch overlaid for 4 tones

3.4.3.2 CLASSIFICATION EXPERIMENTS

For these experiments, segments were extracted from the data base using the supplied labels for vowels. Segment durations were varied from 100 ms to 1000 ms in steps of 100 ms. For each segment length, DCTC and/or pitch was computed and then represented with DCSC terms. Based on results of pilot experiments, the DCTCs were only computed from the low frequency spectra (75Hz to 800 Hz). Three conditions were tested:

DCTCs + Pitch: 5 DCTCs encoded with 7 DCSC terms each, pitch encoded with 7 DCSC terms (42 total features)

Pitch only: pitch feature encoded with 7 DCSC terms (7 features in total)

DCTCs only: 6 DCTCs encoded with 7 DCSC terms each, resulting in a total of 42 features.

Each of the above conditions were also repeated for 5 time “warping” factors (0, 5, 10, 15, 20). These time warping factors control the effective length of the time window, in conjunction with the total segment length. Basically, these curves represent the time window used for each sliding block of frame-based features used to compute DCSCs. These warping factors are illustrated in Figure 12, in terms of their effect on the first basis vector in the DCSC calculations. For warping of 0, the first DCSC basis vector is a constant. As the warping increases, the first DCSC basis vector becomes progressively more peaked and narrow, thus reducing the effective time duration of the basis vectors computed with that degree of warping.

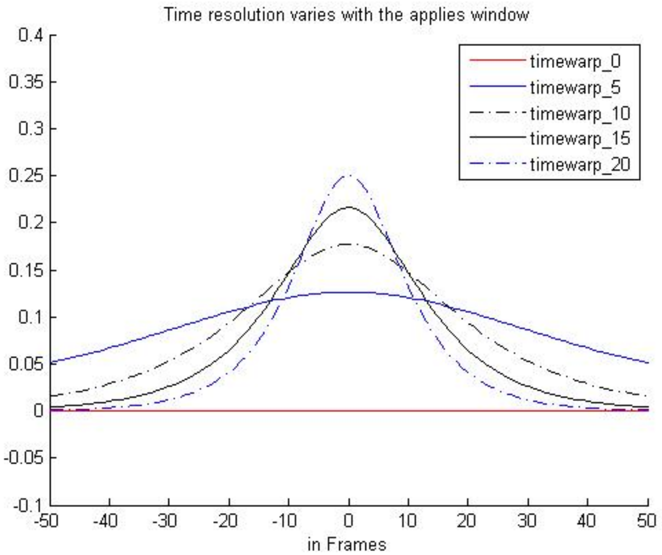


Figure 12: Time warping curves for tone classification experiments

These features were classified with a neural network classifier having two hidden layers (50 hidden nodes, 25 hidden nodes) and an output layer of 4 nodes. Test results, for each of the three features sets, as a function of segment length are shown in the following figures

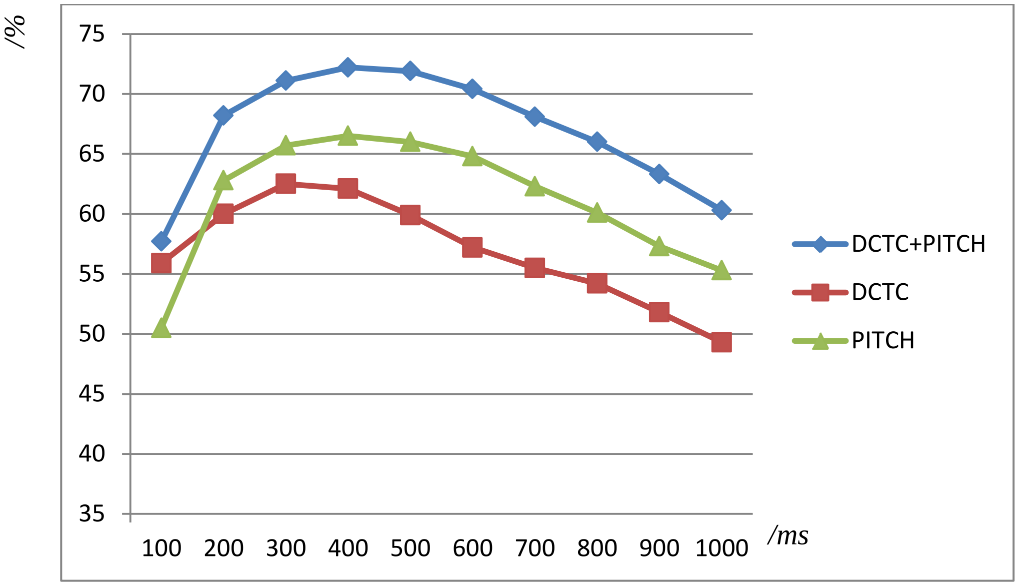


Figure 13: Tone classification accuracy for three features sets as a function of total segment length, using time warp = 0.

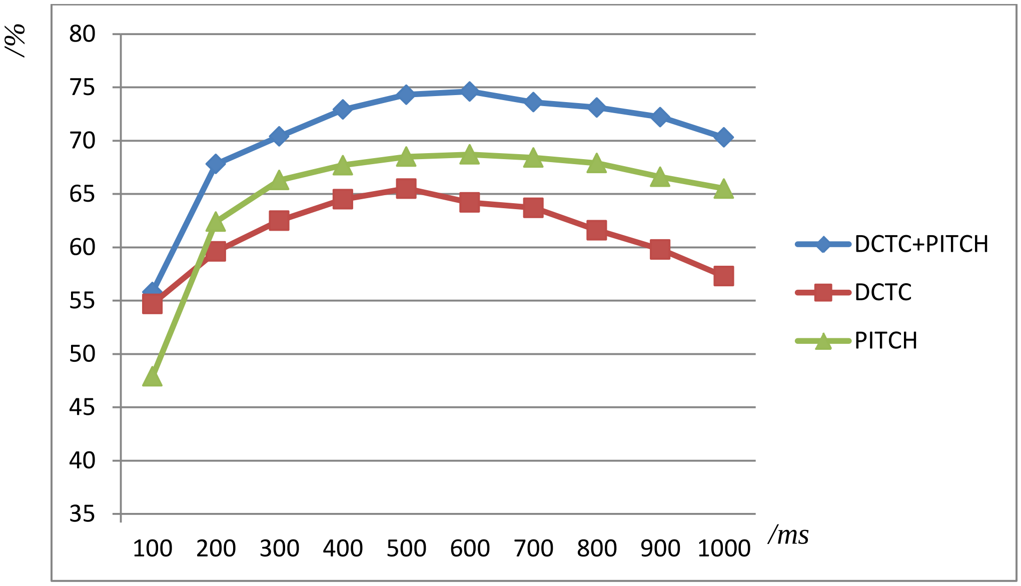


Figure 14: Tone classification accuracy for three features sets as a function of total segment length, using time warp = 5

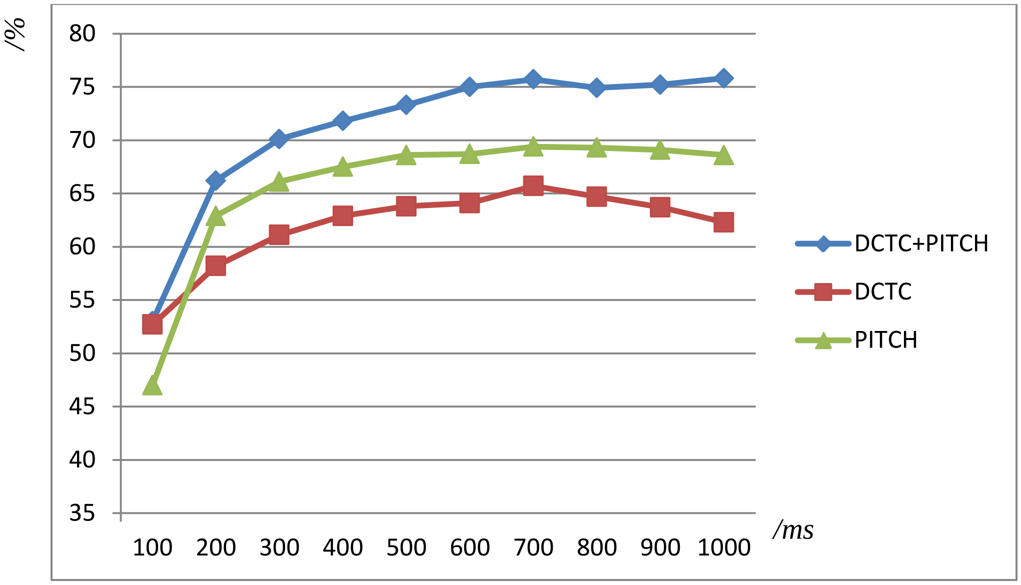


Figure 15: Tone classification accuracy for three features sets as a function of total segment length, using time warp = 10.

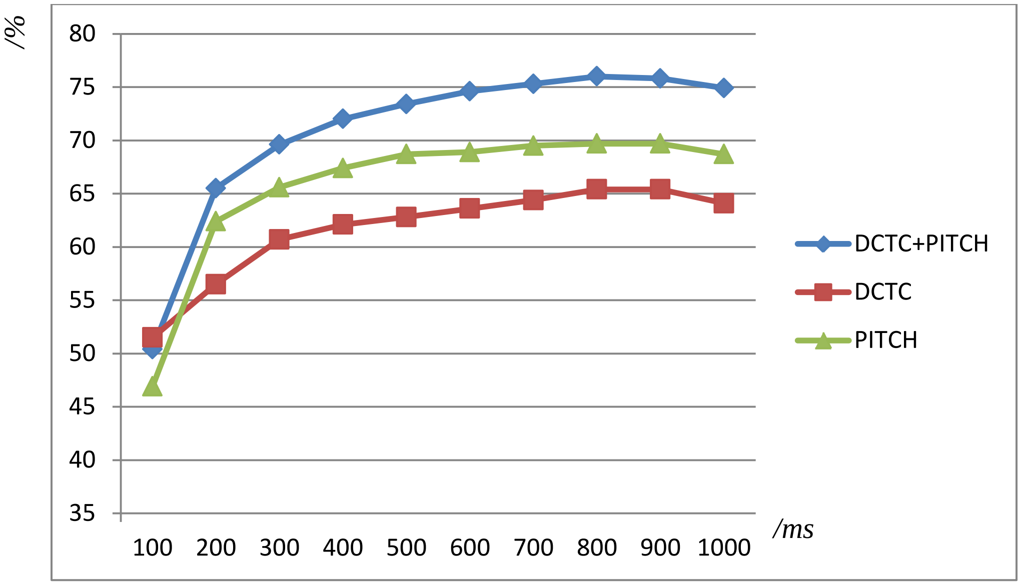


Figure 16: Tone classification accuracy for three features sets as a function of total segment length, using time warp = 15.

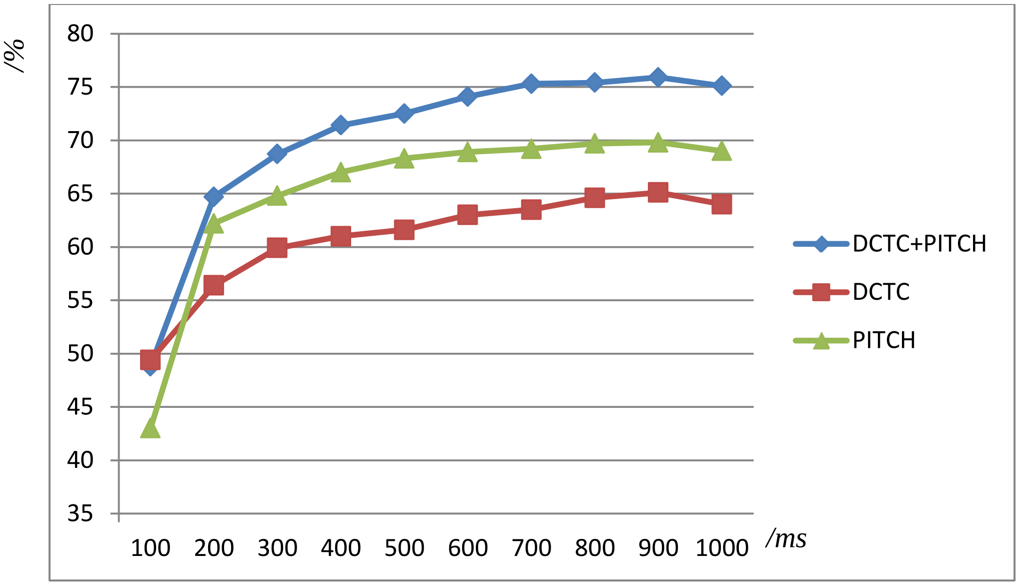


Figure 17: Tone classification accuracy for three features sets as a function of total segment length, using time warp = 20.

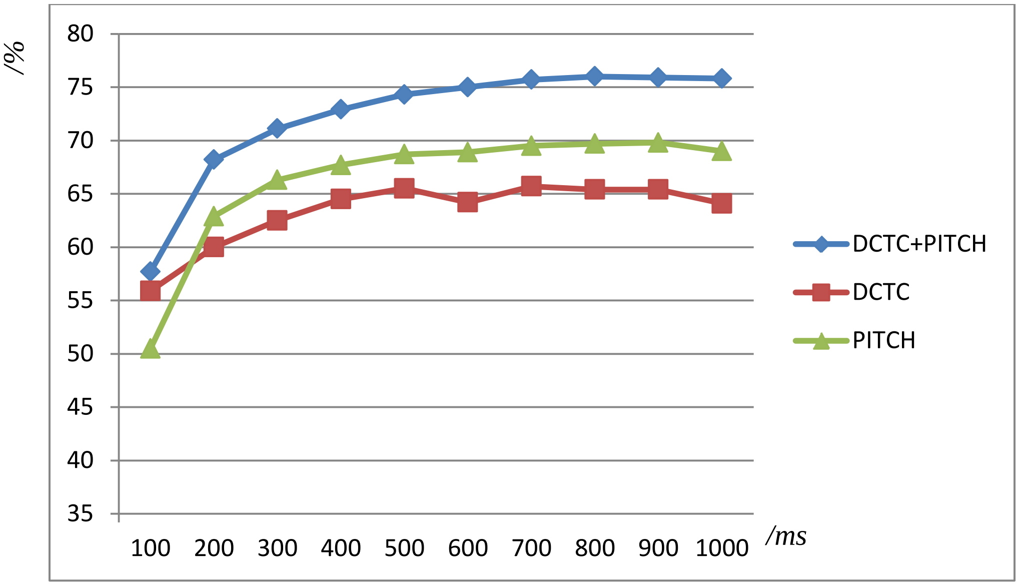


Figure 18: Tone classification accuracy for three features sets as a function of total segment length, using the best warping value for each case.

3.4.3.3 ANALYSIS OF RESULTS

The results of the automatic tone classification using a neural network classifier can be summarized as follows;

1. The overall highest tone classification accuracy is approximately 76% using a combination of spectral/temporal features (DCTCs/DCSCs) and a pitch contour. These best results were obtained using a low frequency range (75 Hz-800Hz) and longtime interval (800ms).
2. The highest tone classification accuracy using only a pitch contour is approximately 70%, using 7 DCSC terms to represent pitch over an interval of at least 800 ms.
3. The highest tone classification accuracy using spectral/temporal features only is approximately 65%, with spectral/temporal features computed from 75-800 Hz frequency range and 800 ms or longer frequency range.

3.4.3.4 DISCUSSION AND CONCLUSIONS

These automatic results compare favorably to the results obtained by native Mandarin human listeners. For most cases tested with listeners, (especially the vowel only and single syllable cases), the human accuracy was lower than the 76% accuracy reported for the automatic method. It is interesting that fairly high tone classification accuracy can be obtained without using pitch—approximately 65% using spectral/temporal features only. Although the pitch contour appears to be the best single feature for tone classification, it is better yet to augment pitch with spectral/temporal features. The best results were obtained with nearly a 1 second segment interval; however, due to the time warping factor used (see Figure 12), the effective best interval was closer to 400 ms.

We hypothesize that higher accuracy could be obtained by increasing the size of the training database (training results, not reported, were considerably higher than the reported test results), and/or normalizing the pitch contours by average values from each speaker.

3.4.4 Tone Recognition Using HMMs

3.4.4.1 INTRODUCTION

Research reported in section 6.4 is also for tone recognition, using completely automatic methods, based on HMMs. The section begins with a brief summary of some labeling modifications that were needed in order to be able to do some of the experiments reported in this section.

In the RASC863 database tones are labeled as numbers after syllable finals. The numbers 1 to 4 represent high tone, rising tone, dipping tone and falling tone, respectively. Syllable finals with a neutral tone, which indicates that there is no major change in pitch and the pitch is not particularly high, is labeled with number 0. Their equivalent labels used for HTK experiment are “H,” “R,” “D,” “F” and “N.”

Also, all consonants (syllable initials which have no tone) will be modeled as one "INIT" model as the goal here is recognizing tones (explicit tone recognition0. "SIL" is used for modeling all silence and short pause intervals. Therefore, a total of 7 HMMs, as shown in Table 14, are created and trained and evaluated through the experiments.

Table 14: HMMs required for tone recognition (monotone)

Tones (5)	**Other (2)**
High (H)	Silence (SIL) Syllable initials (INIT)
Rising (R)	
Dipping (D)	
Falling (F)	
No-tone (N)	

3.4.4.2 TONE RECOGNITION USING HIDDEN MARKOV MODELS

In contrast to the tone classification work reported in the previous section, in this section tone recognition work is reported. Note that for classification, labeled boundaries are used. For the more difficult recognition task, boundaries must be automatically determined. As for all recognition results given in this report, the automatic recognizer is built with Hidden Markov Models (HMMs) using HTK Ver3.4. This toolkit allows complete flexibility in terms of the number of mixtures, number of states, types of transitions, and provides for language modeling. Note that the recognizer is expected to be of lower accuracy than the classifier, both because recognition is more difficult than classification, and since more categories are modeled (7 vs 4 for the classification task). On the other hand, as mentioned later, the tone recognizer used a bigram language model, versus no language model for the tone classification, thus giving additional information to the recognizer, making the task easier.

5 tones (High, Rising, Dipping, Falling and Neutral), silence, and all syllable initials (consonants) were modeled by 7 left-to-right 3-state HMMs with no skip states allowed. A bigram language model was also built from the statistical data of all transcriptions.

Figure 19 shows the results for the experiment that compares two feature sets; pitch plus 47 DCTC/DCSCs, and DCTC/DCSCs only. The total number of features used for this experiment was fixed at 48. The frequency range used for DCTC calculations was 50 to 7000 Hz and 16 mixtures were used for HMM training. The time warp was fixed at 21 for all these experiments.

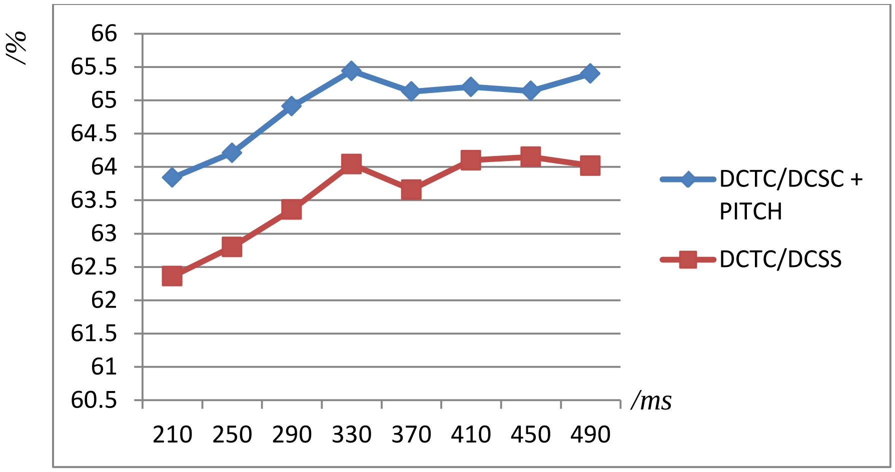


Figure 19: Tone recognition accuracy for two features sets as a function of total segment length

A series of test similar to those described in section 6.3.2 for investigating the optimal time warp was performed here. 11 DCTCs computed from a frequency range of 50Hz to 5500Hz were each represented by 4 DCSCs. Log energy and 3 DCSC of pitch were also used as the features. Thus, in total 48 features we used. Typical numbers used in research for tone recognition (200ms, 250ms and 300ms) were chosen as the block length for computing DCSCs. Then a range of time warping values (from previous research) was tested for each time length. The best accuracy of 66.2% was obtained using 200ms block length with time warp 13 (Figure 20). 16 mixtures were used to model 7 HMMs (Table 14) in this test.

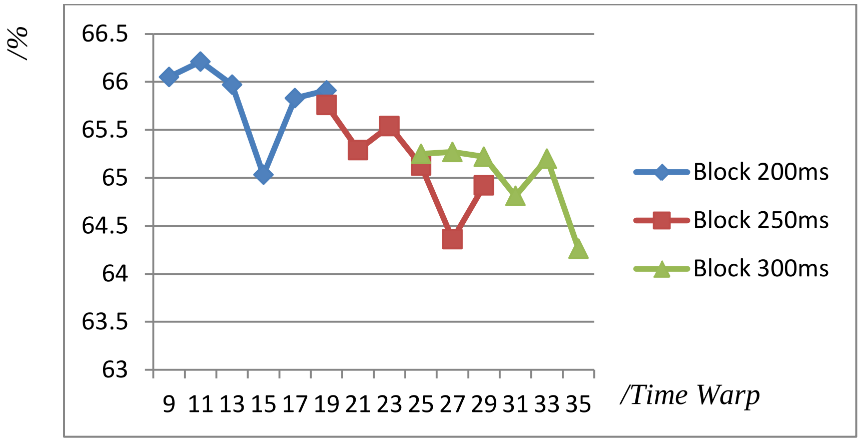


Figure 20: Tone recognition accuracy for three block lengths as a function of time warp value

The results of searching for the optimal segment length and time warping value that were used for computing DCSC terms of pitch are shown in Figure 21.

In this experiment, the components of the feature set, number of mixtures used for GMMs, frequency range of DCTC compression, etc., are all identical to the values used for the results shown Figure 20. The block length for the temporal trajectory features of DCTCs was fixed at 200 ms (that is a 200ms interval was used to compute the DCSCs of the DTCTCs), and the time warping factor was selected as 13, as used before. However, for the case of pitch, various block lengths and time warping factors were explored for the DCSC expansion of pitch.

Results are shown in Figure 21. Note, for these results, the 44 DCTC/DCSC terms and log energy were computed identically as for the results given above. However, the way in which the pitch feature (which is likely to be the single most important feature), is varied, in terms of block length and time warping factor used for the DCSCs of pitch. It appears that the best choice for pitch representing pitch trajectories is to use a block length of about 110ms and time warping of about 0.

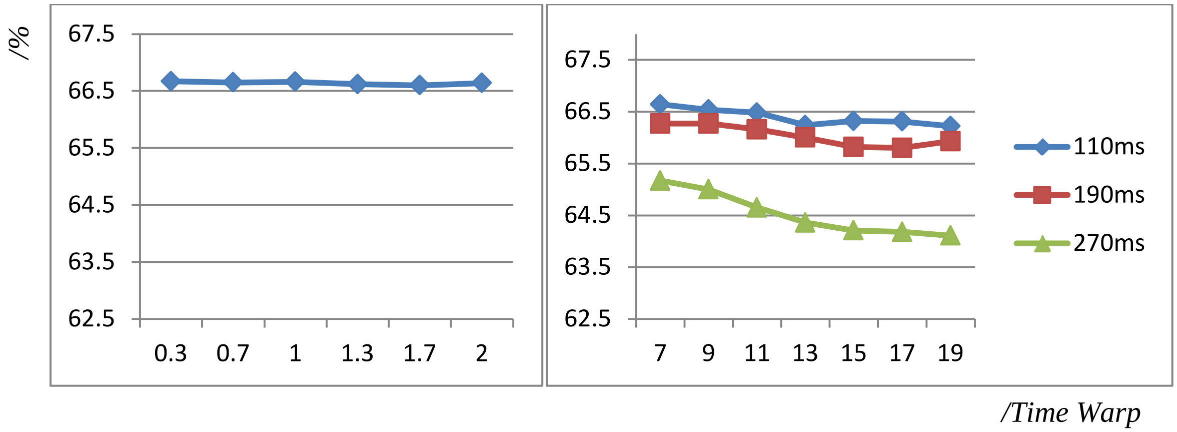


Figure 21: Tone recognition accuracy for three block lengths as a function of time warp value for pitch DCSCs

Another way to improve the overall accuracy is employing more mixtures for HMM training. The data amount given from RASC863 is only able to support 24 mixtures for Gaussian modeling of 60 basic phones (e.g. running a basic phone recognition test in Figure 6 without crashing for the error of “Too few observations”). Due to the fact there are only 7 HMMs required to be modeled by the exact same data for the tone recognition job, the number of mixtures used on each tone (or initial or silence) can stably go up to 65 (or more).

Overall, an accuracy of 71.4% tone recognition accuracy has now been achieved (by January of 2013) using 12DCTCs/5DCSCs, 5 DCSCs of pitch, and log energy (which actually replaces the last DCSC of the last DCTC). This system of 65 Features was modeled by 65 Gaussian mixtures. The segment block length of the DCSC was set to 200ms with a time warping value of 13 while the segment length of 90ms with no warping was used to compute pitch DCSCs.

3.4.4.3 BITONE RECOGNITION USING HMMS

Two methods for splitting and regrouping tone labels were introduced (Figure 3 and Figure 4). The hypothesis was that tones might be more accurately recognized in pairs rather than in isolation. Method 1 uses modeling for 5 single tones (High (H), Rising(R), Dipping(D), Falling(F), Neutral(N)) and all their combinations except NN (29 total) , plus one for silence (SIL). Thus there are 30 models total for method 1. The models in method 2 include all pair-wise combinations of (SIL, H, R, D, F, N) except NN and SN. Thus, for method 2, there are 34 total models. See table 15 for details of the number and names of models.

Table 15: HMMs required for tone recognition (Bitone)

Method 1 (30 HMMs)	**Method 2 (34 HMMs)**
SIL, H, R, D, F, N, HH, HR, HD, HF, HN, RH, RR, RD, RF, RN, DH, DR, DD, DF, DN, FH, FR, FD, FF, FN, NH, NR, ND, NF	SH, SR, SD, SF, HH, HR, HD, HF, HN, RH, RR, RD, RF, RN, DH, DR, DD, DF, DN, FH, FR, FD, FF, FN, NH, NR, ND, NF, HS, RS, DS, FS, NS

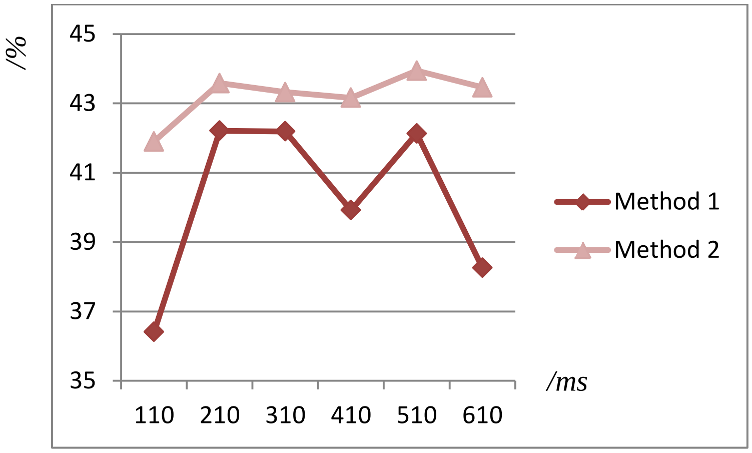


Figure 22: Tone recognition using Bitone structure

Figure 22 shows the results obtained for recognizing Bitone models from two methods in table 15. 24 DCTC/DCSC features were selected and modeled by 8 Gaussian mixtures (same configuration as used for results shown in Figure 19). The time warp value for each DCSC length was chosen to be the optimal for that case. The overall results shown in this figure suggest using a longer DCS (500ms or longer) might be better to encode the sound having a bitone structure for it very likely longer than monotone vowels.

Overall, the Bitone method did not appear promising and is not being actively pursued at this time. The major concern about this technique is that, compared to only 7 models needed for monotone modeling, the Bitone methods require more than 30 HMMs, and the amount of training data is the same for both methods. Without a larger database, it would be very difficult to achieve competitive accuracy. Additionally, the DCSC features which are based on long time intervals, likely account for most of the coarticulation effects that Bitones are intended to compensate for.

3.4.5 Conclusions

In this section, we looked at the ability of human listeners to recognize tones extracted from continuous Mandarin Chinese, we looked at the accuracy of two automatic methods for classifying (using neural networks) or recognizing (using Hidden Markov Models), and compared the human and automatic methods.

The experimental results obtained with native speakers indicate that humans need context to recognize tones very accurately. Nonnative speakers find this task nearly impossible. Without context, machine recognition and human recognition have about the same accuracy, but with different patterns of errors.

The most interesting and potentially significant result of this work is that reasonably accurate tone classification and recognition can be obtained without using a pitch feature. However, pitch does seem to be the single most important feature for tone recognition, and adding pitch to spectral/temporal features for tone recognition is preferred.

3.5 Spectral Analysis Methods

3.5.1 Introduction

All automatic speech recognizers perform spectral analysis at the front end which converts the speech signal, possibly noisy and/or degraded, into values from which useful features can easily be computed. The front end spectral analysis is performed by calculating the short time Fourier transform (STFT) of the speech signal, either using an FFT, a filter bank, or a combination of the two methods. For the combination method, the filter bank is approximated by summing weighted combinations of FFT magnitude values. The filter bank approach, even if derived from FFT values, is thought to be advantageous since it can be designed to mimic the functionality of the cochlea of the human auditory system, such as a nonlinear (“warped”) frequency scale.

The majority of ASR systems are implemented using a Mel filter bank as the spectral analysis front end, followed by a cosine transform based feature extraction which is shown to outperform other signal processing methods [23]. Very recently, there is another filter bank that has been presented as a superior alternative to the triangular-shaped Mel filters called the Gammatone filter bank, which simulates the motion of the basilar membrane within the cochlea of the human auditory system. It was first introduced by Johannsma (1972) to describe the shape of the impulse response function of the auditory system as estimated by the reverse correlation function of neural firing times. The general thinking is that since the Gammatone filter bank approximates the human auditory system better than the Mel filter bank, it should also be superior for ASR applications.

The Gammatone filter is defined in the time domain (impulse response function) as:

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi ft + \emptyset) \qquad (1)$$

Where f is the frequency, $\emptyset$ is the phase of the carrier, a is the amplitude, n is the filter order, b is the bandwidth and t is time.

Front-end spectral analysis can also be performed without using any filter bank, but simply using an FFT directly. In either case, spectral values (that is FFT values or filter bank outputs, both converted to magnitudes), are typically reduced in dimensionality using some type of cosine transform. If the filter bank step is used, cosine basis vectors can be used directly. However, if the FFT magnitudes are used as the direct input to the cosine transform, the cosine basis vectors should be modified to account for the non-uniform frequency resolution. In order to incorporate spectral trajectory information into ASR feature sets, additional terms are generally computed from blocks of frame-based features, such as delta terms.

In the following sections we compare spectral features computed as cosine transforms of filter bank outputs with features computed as modified cosine transforms (DCTCs) of FFT spectral magnitudes directly. We also compare delta type trajectory features with trajectory features computed over much longer time intervals using another set of modified cosine basis vectors (DCSCs). More details of the more common spectral and feature calculation method (MFCCs with delta and delta-delta terms are given in [24] and [25]. More details of the DCTC/DCSC general method are given in [26] and [27]). All the methods are evaluated using as much similarity of parameters and recognizer as feasible (such as frequency range, # of HMM mixtures, etc.) in order to make comparisons most meaningful.

3.5.2 FFT Based Spectral Analysis

The most common front end spectral analysis method for speech recognition uses a frame-based approach in which the time varying speech signal is described by a stream of feature vectors, with each vector reflecting the spectral magnitude properties of a relatively short (10-25ms) segment (frame) of the signal. For experimental results reported in this paper, 16 kHz sampling rate speech signals are short-time Fourier transform (STFT) analyzed using a 10ms Kaiser window with a frame space of 2ms. The spectrogram of a typical speech signal is as shown in Figure 23. The FFT spectral values are used as the front-end for DCTC/DCSC feature extraction, as described later. The frame length and frame spacing mentioned were empirically determined as providing most accurate ASR results.

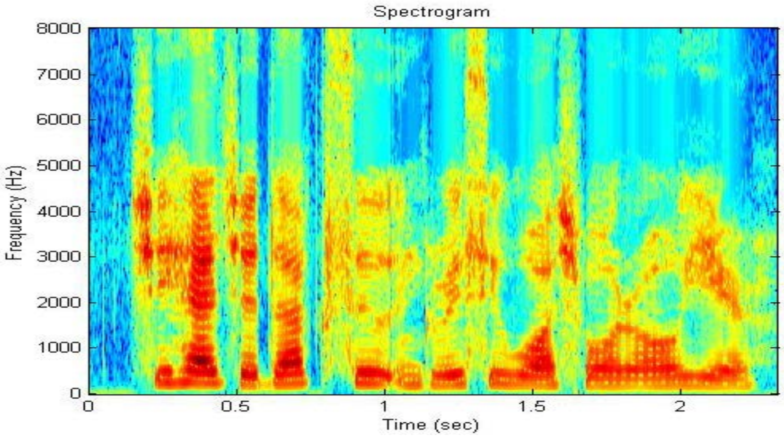


Figure 23: FFT spectrogram

3.5.3 Filter Bank Based Spectral Analysis

A filter bank can be regarded as a crude model for the initial stages of transduction in the human auditory system. A set of band pass filters is designed so that a desired portion of the speech band is entirely covered by the combined pass bands of the filters composing the filter bank. The output of the band pass filters are considered to be the time varying spectrum representation of the speech signal.

For the experiments given in this paper, we evaluate two commonly used filter banks: the Mel filter bank and Gammatone filter bank. Either the DCTC/DCSC method (but without frequency warping) or the more common method used for MFCC features (i.e., delta terms rather than DCSCs) are used. Results are compared for the filter bank approaches versus the FFT-only spectral method.

3.5.3.1 MEL FILTER BANK

The Mel filter bank is series of triangular band pass filters, as depicted in Figure 24, designed to simulate the band pass filtering believed to be similar to that occurring in the auditory system.

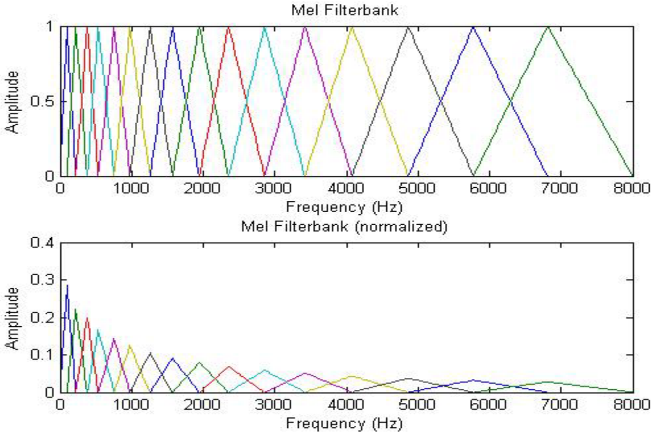


Figure 24: Frequency response of 16 channel Mel filter bank and the normalized versions of the filters, as used for MFCCs.

To convert the frequency in Hz into frequency in Mels the following equation is used:

$$m = 1127.01048 * log_e\left(1 + \frac{f}{700}\right) \qquad (2)$$

On a linear frequency scale, the filter spacing is approximately linear up to 1000 Hz and approximately logarithmic at higher frequencies. For actual implementation, the Mel filter bank is computed by first computing the power spectrum with an FFT, and then multiplying the power spectrum by the Mel filter bank coefficients. In Figure 25 is shown a spectrogram based on 32 Mel filters. Note that this spectrogram is qualitatively similar to the direct FFT spectrogram shown in Figure 23. The details of the two spectrograms are quite different since the frequency range is more quantized in Figure 25 and the frequency scale is effectively in Mels rather than linear. However, it should be noted that the Mel spectrogram, or Mel filters, are derived from the FFT spectral values and thus are simply an intermediate step in processing.

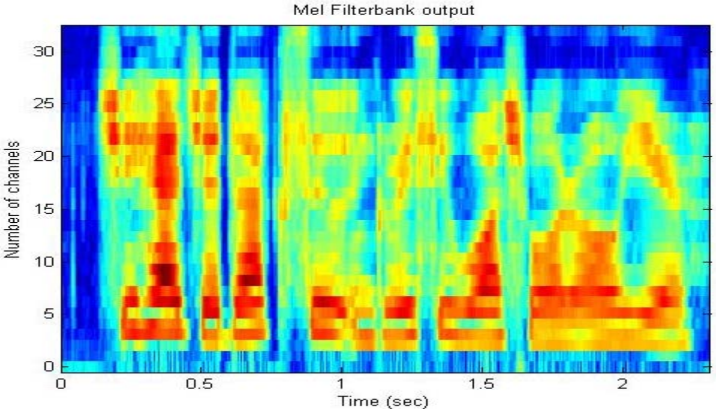


Figure 25: 32 channel Mel Spectrogram

3.5.3.2 GAMMATONE FILTER BANK

A Gammatone filter is a linear filter with impulse response described as the product of a (gamma) distribution and sinusoidal (tone), hence the name Gammatone. The filter bank is a combination of individual Gammatone filters with varying bandwidth based on the Equivalent Rectangular Bandwidth (ERB) scale. For moderate sound pressure levels, Moore et al [28] [29] estimated the size of ERBs for humans as:

$$ERB[f] = 24.7 + 0.108 * f_c \qquad (3)$$

The value ERB[f] is used as the unit of center frequency f_c on the ERB scale. For example, the value of ERB[f] for a center frequency of 1 kHz is about 132.64, so an increase in frequency from 975 to 985 Hz represents a step of one ERB[f]. Each step in ERB roughly corresponds to a constant distance of about 0.89 mm on the basilar membrane [30].

As the center frequency increases the bandwidth of the filter bank increases. A 16 channel Gammatone FFT based filter bank frequency response is shown in Figure 26.

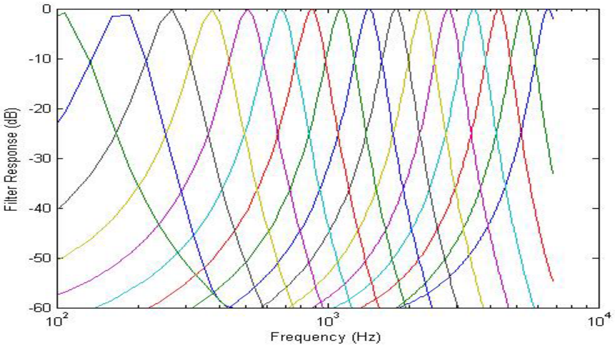


Figure 26: Frequency response of 16 channel Gammatone filter bank

The Gammatone filter bank can be implemented using sums of weighted FFT power spectrum values, exactly as for the Mel filter bank except using the weights corresponding to Figure 26, rather than the Mel filter weights shown in Figure 24. Alternatively, the Gammatone real filters can be implemented as actual IIR or FIR filters, followed by rectification and low pass filters, as depicted in Figure 27. Figure 28 depicts the Gammatone spectrogram of the same sentence as was used to construct the spectrograms for Figures 23 and 25.

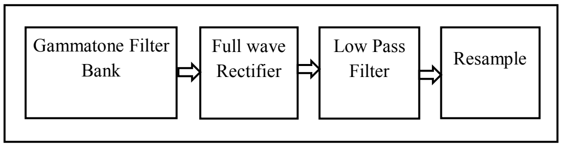


Figure 27: Block Diagram of Gammatone using actual filter (difference equation in first block

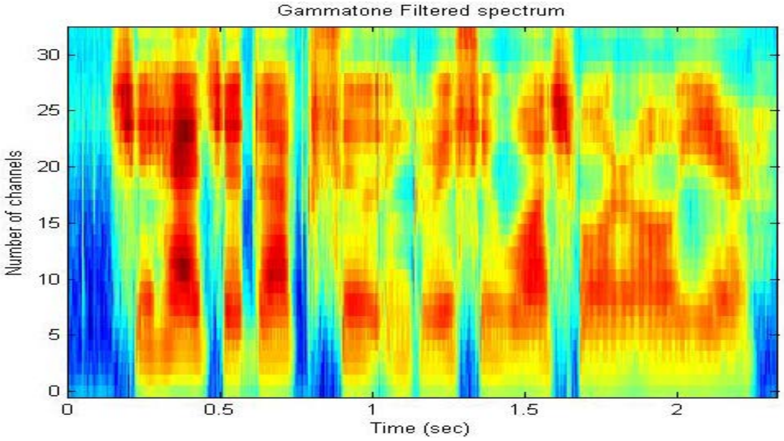


Figure 28: 32 channel Gammatone spectrogram

3.5.4 Experimental Evaluation

Phonetic recognition experiments were conducted using the TIMIT phonetically-labeled database. 3296 sentences from 462 speakers were used for training and 1344 sentences from 168 speakers were used for test. SA sentences were excluded. A frequency range of 100 to 8000 Hz was used for all experiments. Experiments were conducted with clean, 20 dB SNR, 10 dB SNR, and 0dB SNR speech. For all conditions, training and test conditions were matched with respect to noise; additive white Gaussian noise was used for noise.

The objective of the experiments was to compare phoneme recognition accuracy of four spectral analysis methods, as depicted in Figure 29, and also to compare to a control case (13 MFCCs with delta and acceleration terms, or 39 total terms, derived from a Mel filter bank, as implemented in HTK).

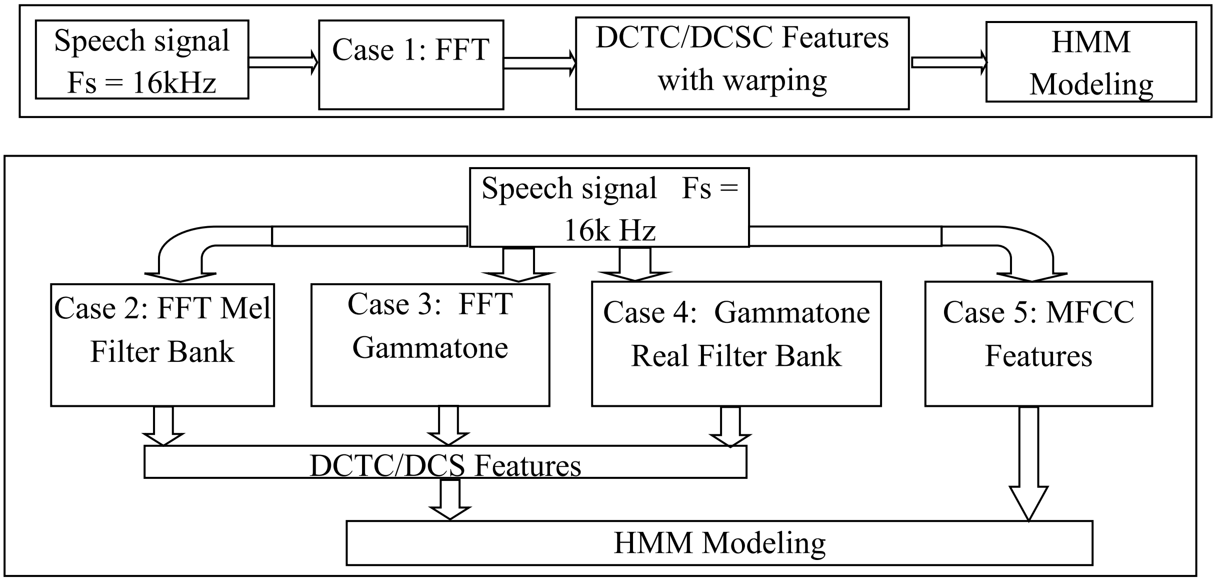


Figure 29: Block diagram of automatic speech recognition process

Five cases, as depicted in Figure 29, and outlined below were tested.

Case 1: FFT spectrum directly used as front end for DCTC/DCSC feature, using frequency warping shown in the Figure 24.

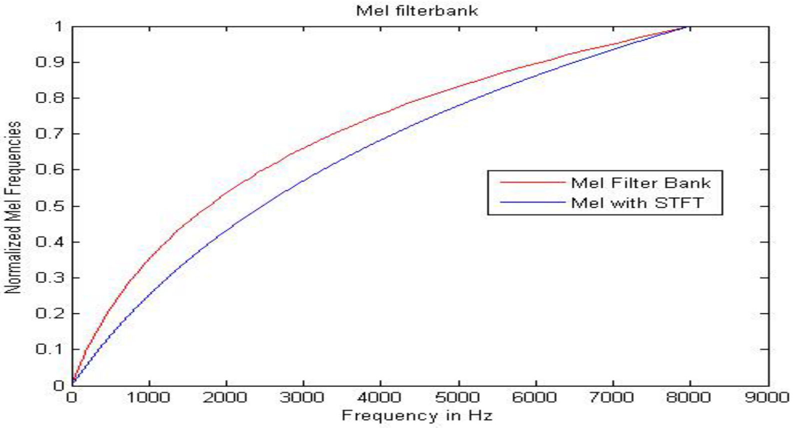


Figure 30: Mel frequency warping used for Mel filter bank center frequencies (top red curve), and "optimum" Mel frequency warping used for FFT-only/DCTC/DCSC method (bottom blue curve)

Case 2: DCTC/DCSC feature extraction applied to Mel filter bank spectrum. Since the filter bank already has warping in it, the DCTC basis vectors have no warping.

Case 3&4: Gammatone filter banks (FFT-based and actual filters cases) used as front end for DCTC/DCSC features, with no frequency warping used for DCTCs.

Case 5: HTK MFCC features with delta and acceleration terms.

For all experiments with DCTC/DCSC features, a frame spacing of 2 ms (500 frames per second) was used. Blocks were comprised of 150 frames (300 ms) and spaced 8 ms apart (125 blocks per second). Experiments were conducted with both 78 features (13 DCTCs times 6 DCSCs), and the more standard 39 features (13 DCTCs times 3 DCSCs).

HMMs with 3 hidden states from left to right with 16 Gaussian mixtures were used for phonetic recognition experiments. A total of 48 (eventually reduced to 39 phones) context independent monophone HMMs were created using the HTK toolbox (Ver3.4) [16]. The bigram phone information extracted from the training data was used as the language model.

3.5.5 Results

Phonetic recognition accuracy (based on 39 phones) obtained for all 5 cases is given in Table 16. It can be seen that there is negligible or no improvement when filter bank techniques are used. For results in Table 16, 39 features were used. The experiment was repeated with 78 features for all cases except MFCC, and results are given in Table 17.

Table 16: Accuracy (%) comparison for 39 features

SNR (dB)	FFT only	Mel FB	Gammatone FFT FB	Gammatone Real FB	MFCC
Clean	**69.18**	**68.52**	**69.75**	**69.12**	**62.83**
20 dB	**64.19**	**63.47**	**63.74**	**63.39**	
10 dB	**56.32**	**54.99**	**55.76**	**55.03**	
0 dB	**42.21**	**41.45**	**41.42**	**40.53**	

Table 17: Accuracy (%) comparison for 78 features

SNR (dB)	FFT only	Mel FB	Gammatone FFT FB	Gammatone Real FB
Clean	**71.20**	**69.70**	**71.06**	**70.11**
20 dB	**65.83**	**64.65**	**65.80**	**64.93**
10 dB	**58.03**	**57.01**	**58.13**	**56.94**
0 dB	**43.40**	**42.46**	**42.83**	**41.82**

Table 16 and 17 show the experimental results for 16 mixtures. The same experiments were performed, except 32 mixtures were used for HMMs, and the results are as shown in Table 18 and 19.

Table 18: Accuracy (%) comparison for 39 features

SNR (dB)	FFT only	Mel FB	Gammatone FFT	Gammatone Real
Clean	**71.10**	**69.71**	**71.04**	**70.97**
20 dB	**65.74**	**65.41**	**65.47**	**65.19**
10 dB	**57.04**	**56.64**	**56.45**	**56.25**

Table 19: Accuracy (%) comparison for 78 features

SNR (dB)	STFT	Mel FB	Gammatone FFT	Gammatone Real
Clean	**73.00**	**71.10**	**72.99**	**72.23**
20 dB	**67.89**	**67.05**	**67.69**	**66.95**
10 dB	**59.43**	**58.68**	**59.39**	**58.96**

As yet another test, Table 20 shows the accuracy obtained with the Gammatone filter bank as the number of channels is varied from 8 to 128. Although there is a very slight improvement when using 64 channels, this comes at the expense of more computational time and complexity, so we considered the “standard” as 32 channels for the Gammatone filters. Note that performance for the Gammatone case also decreases slightly if 128 filters are used. It should also be pointed out that, for the Mel filter bank approach, 32 filters were also used for the results given above.

Table 20: FFT Gammatone performance as number of filters is varied.

SNR (dB)/Channels	8	16	32	48	64	128
Clean	**64.51**	**69.36**	**71.06**	**71.29**	**71.37**	**71.05**
20 dB	**60.06**	**64.78**	**65.80**	**65.82**	**65.94**	**65.90**
10 dB	**50.78**	**56.03**	**58.13**	**58.12**	**59.33**	**58.06**

More experiments were performed using Mel filter bank by computing the logarithm before applying the triangular filters as stated in [33], or computing the logarithm after the filter bank output as stated in [34], and as used for the Mel filter bank results reported above. Results are shown in Tables 21 and 22. For ease of comparison, the results for log after the filter bank are repeated. Accuracy is higher for cases where the log is after the filter bank, as was used for the results reported above.

Table 21: Accuracy (%) of Mel FB for 39 features and 16 mixtures

SNR (dB)	Mel FB (with log before FB)	Mel FB (with log after FB)
Clean	**67.87**	**69.71**
20 dB	**62.83**	**65.41**
10 dB	**54.10**	**56.64**

Table 22: Accuracy (%) of Mel FB for 78 features and 16 mixtures

SNR (dB)	Mel FB (with log before FB)	Mel FB (with log after FB)
Clean	**69.30**	**71.10**
20 dB	**64.28**	**67.05**
10 dB	**55.83**	**58.68**

3.5.6 Conclusion

From the experimental data, we conclude that FFT-based spectral analysis in both clean and noisy conditions with a Mel-like frequency scale incorporated using frequency warping for DCTC features performs nearly identically to cochlea-motivated filter bank spectral analysis. Directly using the FFT spectrum, without the intermediate filter bank prior to feature calculations, has the advantage of simplicity and would appear to be a better front end strategy for spectral front end calculations for speech processing. The DCSC method for computing spectral trajectory features is experimentally shown to result in much higher ASR accuracy than obtained with delta and delta-delta terms.

3.6 Report on Mandarin Syllable Level Label Aligning

3.6.1 Introduction and Motivation

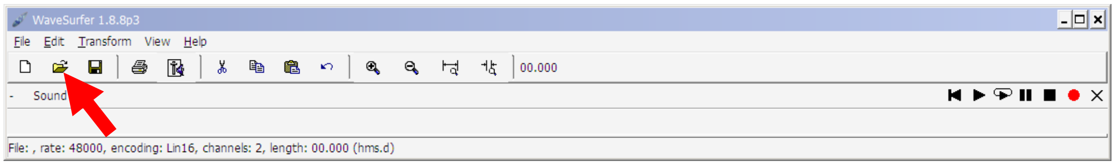
Previously (another report), we reported that automatic methods were used to label Mandarin at the syllable levels. Similar methods were used for detailed labeling of English and found to perform well. However, careful checking of the labels for the Mandarin showed that the syllable level labels were extremely inaccurate (probably of very little value). In order to develop better automatic methods (the real objective), it is very useful to first have a good set of manual labels. Given that several native speakers of Mandarin, all electrical engineering majors, who had come to the US to attend Binghamton University from China in 2011, and were available and working as part of the speech group at SUNY Binghamton, these students were recruited to perform the tedious but important task of accurate manual labeling of the Mandarin database. Five students participated, and each of the students was asked to correct the labels for 60 passages (total of 300 passages). The students began with the automatically determined labels, and using tools and methods described below, listened and observed the speech files to correct these labels (starting and stopping times for each monosyllable). The remainder of this section of the report describes the process the students used for this manual labeling.

3.6.2 Preparation

The aligning work was done with software called "WaveSurfer." Two files were used:

- wavesurfer.exe (available at http://sourceforge.net/projects/wavesurfer/)
- wavesurfer.conf

The configuration file is not necessary to use WaveSurfer. However, it was created by one of the former researchers in our group (Dr. Montri Karnjanadecha) so that WaveSurfer would be more useful as an aligning tool. To begin the labeling process, the researcher should first run "WaveSurfer.exe" and a window, as shown in Figure 31, will be displayed.



File: , rate: 48000, encoding: Lin16, channels: 2, length: 00.000 (hms.d)

Figure 31: WaveSurfer Window

After clicking the "file open" icon (red arrow), the researcher can browse to a "*.wav" file (wave file) and open it. After clicking "Open," the researcher will see a dialog (Figure 32).

The researcher should select "wavesurfer" and click "OK." (When there is no "wavesurfer.conf" in the same directory as "WaveSurfer.exe," the researcher will not be able to choose "wavesurfer" configuration. The researcher can choose "Demonstration" configuration, which is very similar.) The waveform and spectrogram of the "*.wav" file will be displayed on the screen (Figure. 32).

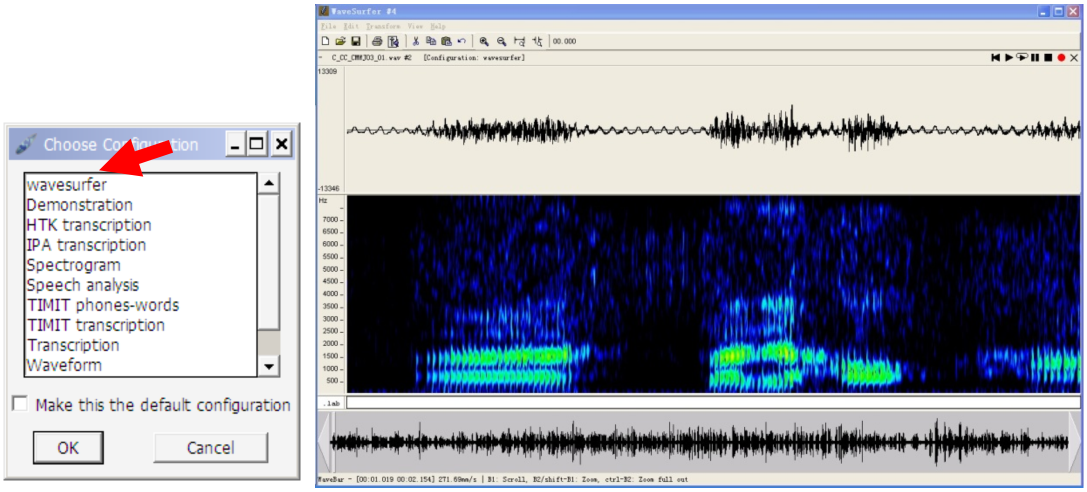


Figure 32: Typical initial waveform and spectrogram display

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

The researcher should right-click anywhere inside the transcription pane (the blank pane inside the red rectangle in Figure 33) and select "Load Transcription…" Then the researcher should browse to the matching "*.lab" file (label file) and click "Open." The labels will then appear (Figure. 33).

Figure 33: Illustration of label display with markers

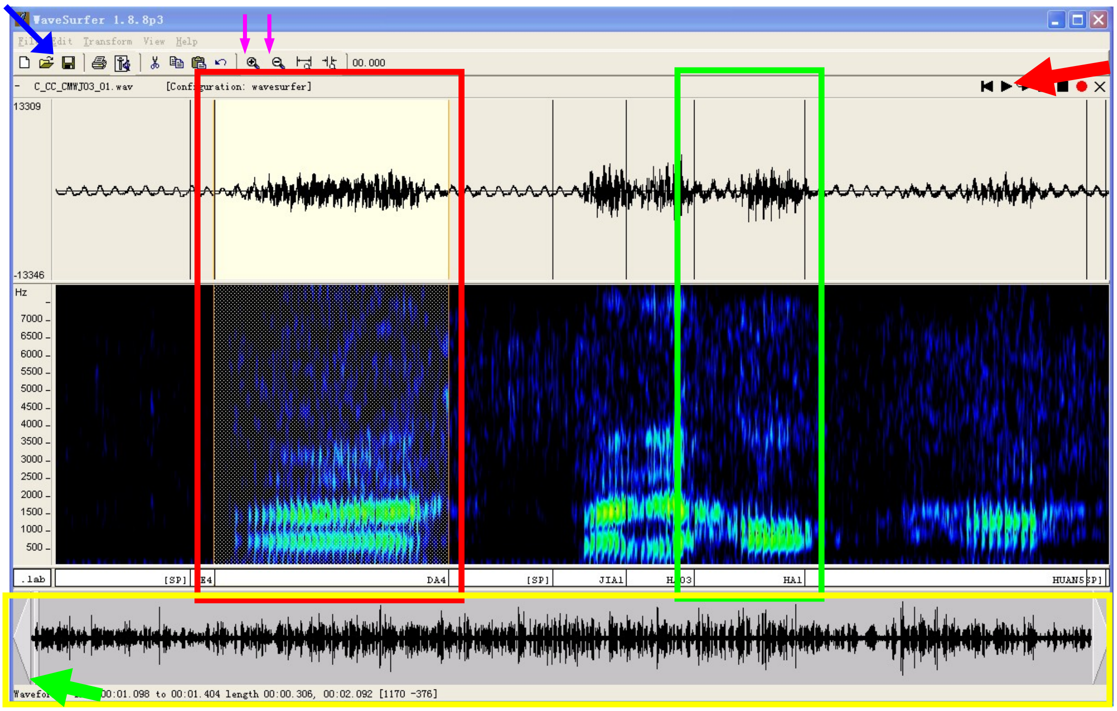
3.6.3 Basic Functions

Playing Audio:

The researcher can click and drag to select a segment of the speech signal (In Figure 34, the dotted area inside the red rectangle is selected.). To play the audio of a selected part, one can click the "Play" button (red arrow) or press the space key. If every label in "Figure 34" is correct, when the "Play" button is pressed, the audio should match the Pinyin pronunciation of "DA4." Similarly, the segment of speech inside the "label borders” of “HA1” (inside the green rectangle, “label borders” refer to the vertical black lines) should match the Pinyin pronunciation of “HA1.”

Inserting, Deleting and Adjusting Labels (details in “Frequent Problems” part of this report). The right-click menu from the transcription pane allows a researcher to insert or delete a label. To adjust the position of a label, one just needs to click and drag the label.

After completing the portion in the current window, the researcher can move to the next part by scrolling the mouse wheel. The gray pane at the bottom (in the yellow rectangle) is a miniature of the waveform of the whole wave file. The bar pointed to by the green arrow is a miniature of the waveform on the top window. The researcher can also click and drag the rectangle to move to other parts quickly. One can zoom in and zoom out (pink arrow on top) to adjust the coverage of the current window.



Waveform 00:01.098 to 00:01.404 length 00:00.306, 00:02.092 [1170 -376]

Figure 34: Illustration of displays with Pinyin labels and markers on basic functions

3.6.4 Frequent Problems

1. Positional Problems

Figure 35: Position adjustment

A label and its corresponding audio do not align. This is the most typical problem—an alignment error. To fix this type of error, the researcher needs to click and drag the label markers to move them to the correct position. Figure 35 is a “close-up” of the transcription pane when a label endpoint is being moved.

2. Incorrect Labels

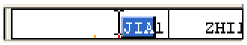


Figure 36: Selected label

A label is incorrect. The researcher can click the word and edit (change the label). Figure 36 is a “close-up” of the transcription pane when a label is being selected.

3. Missing Labels or Redundant Labels

A label is missing or redundant. The researchers can right-click at the label in the transcription pane to insert or delete the label. Figure 37 is a “close-up” of the right-click menu.

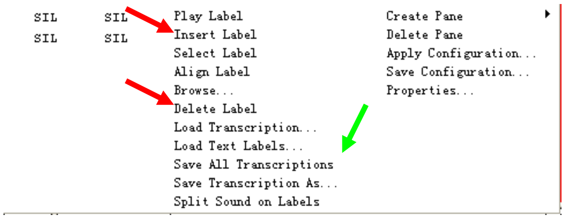


Figure 37: Menu for inserting or deleting labels

4. Lack of Auto-saving

This is an inconvenience rather than a problem due to inaccuracy in the automatic labeling. However, it can cause a large waste of time in the manual labeling process. WaveSurfer does not auto-save the changes in transcription when a computer crashes. Moreover, if the researcher mistakenly clicks “Close,” there will not be any prompt window and the work will be lost. The “Save” button on top left (blue arrow on Figure 34) is not for saving transcriptions. To save the aligning work, one must right-click the transcription pane and click “Save all Transcriptions…” (green arrow on Figure 37). Saving regularly minimizes loss of work.

3.6.5 Feedback from Five Researchers

1. Time Consumption

Table 23: Length of (researcher) time to align files

Time spent on completing one file (hours)	
Maximum	8
Minimum	1
Average	4

2. Improvement on Speed

Generally speaking, all researchers got faster at aligning as they did more files. However, after the researchers completed about 10 files, their speed tended to remain about the same.

3. Accuracy of labeling before Manually Aligning (Approximately)

Table 24: Estimates of accuracy of automatic labeling

Accuracy	Percentage
>50%	3%
40%~50%	8%
30%~40%	8%
20%~30%	25%
10%~20%	33%
<10%	33%

4 Features of the High-Accuracy Files

The files which were more accurate to begin with had the following features (Figure 38 is a snap shot of a high-accuracy file.):

a. low background noise (In Figure 38 , the segment inside red rectangle is without speech. The amplitude of the waveform is almost zero and there is nothing in the spectrogram.)

b. The speaker speaks slowly. (The duration of each syllable is comparatively long.)

c. The speaker speaks clearly (The “pulses” are full and of regular shape.)

d. Generally speaking, female speakers’ speech is aligned better by the automatic aligning.

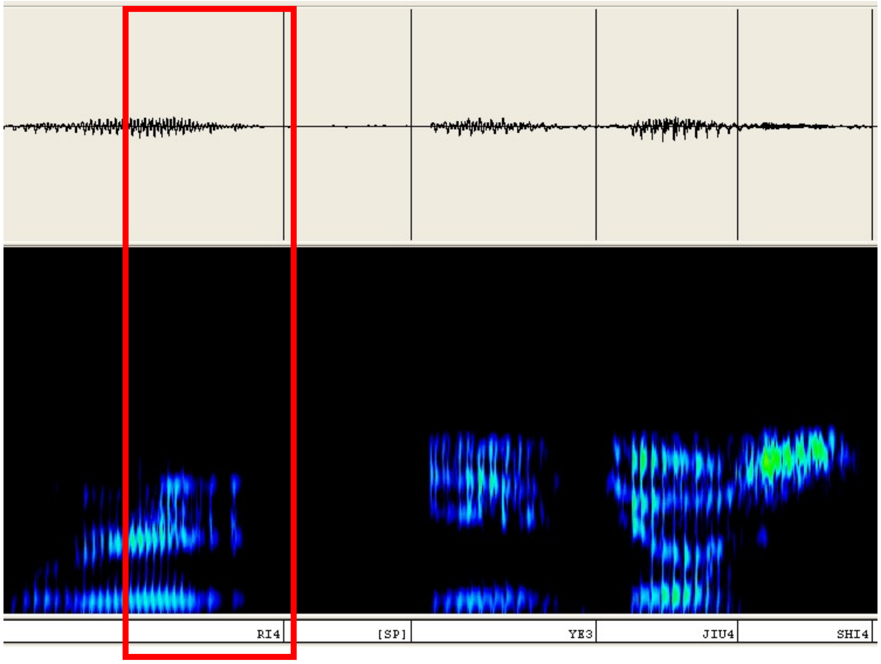


Figure 38: High accuracy example of automatic aligning

5. Features of the Low-Accuracy Files

The files which were less accurate to begin had the following features (Figure 39, Figure 40 and Figure 41 are snap shots of low-accuracy files.):

a. high background noise (In Figure 39, the segment inside the red rectangle is without speech. Its amplitude is almost the same as the amplitude of speech waveform.)

b. The speaker speaks very rapidly. (In Figure 39, the part inside the green rectangle is a segment of fast speech. Before manual alignment, the computer regarded the whole segment as one syllable.)

c. The recorder or microphone is of low quality (In Figure 40, the waveforms are "fluffy." The audio of that speech is "veiled," sounding somewhat like the speaker is behind a curtain. The audio is not very easy to understand.)

d. The speech is too loud. (In many places on Figure. 41, the waveform exceeds the borders of the window. Automatic aligning performed poorly for "loud" files.)

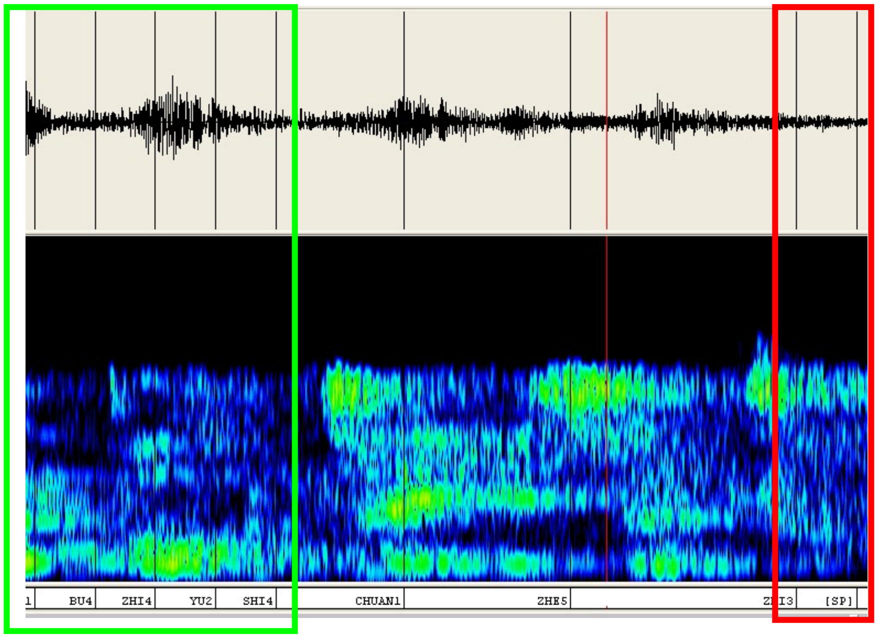


Figure 39: Low accuracy example of automatic aligning (difficult to understand)

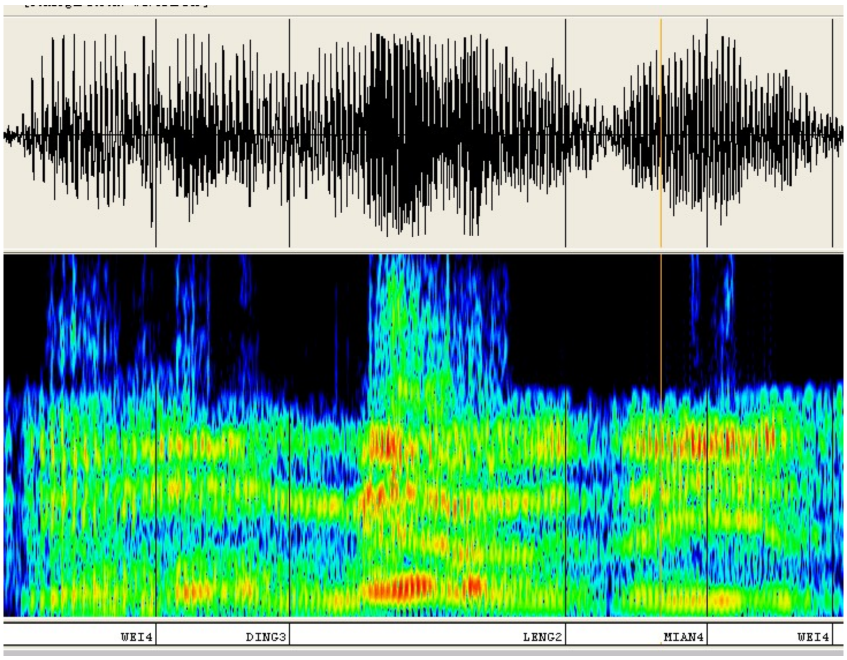


Figure 40: Low accuracy example of automatic aligning (difficult to understand)

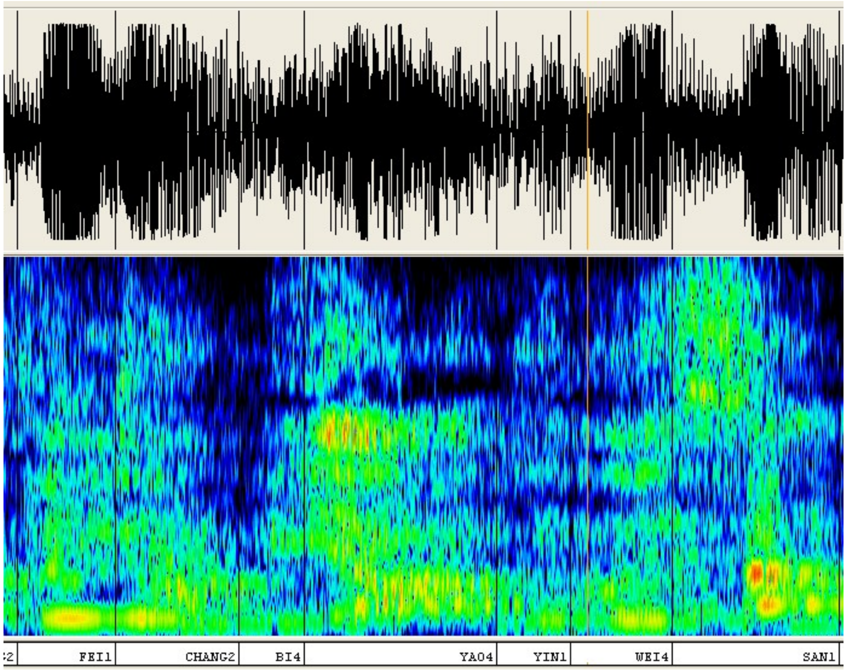


Figure 41: Low accuracy example of automatic aligning (nearly flat envelope)

3.6.6 General Strategy

The manual aligning of a label consists of three steps:

a. Look at the waveform and estimate the general position of a series of syllables. The valleys are very likely to be the boundaries of the syllables. (In Figure 42, the red arrows all point to valleys which were ultimately selected as boundary points.)

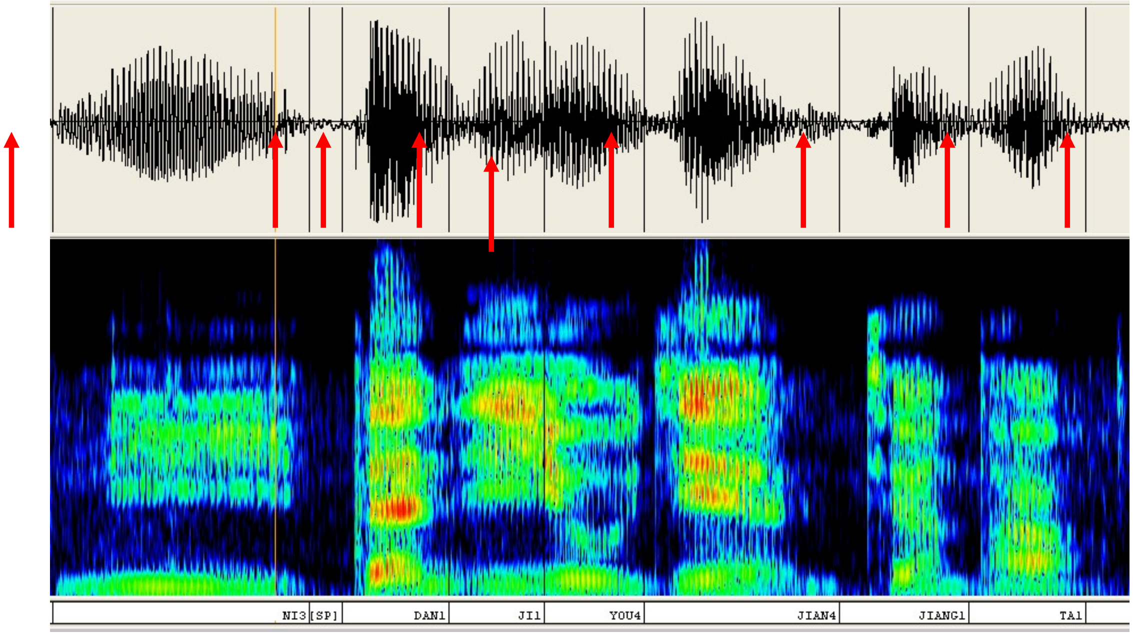


Figure 42: Waveform that is very helpful for estimating syllable position

a. Select a series of syllables and play them. By listening to the audio, find the stopping point of the right-most syllable. (The five researchers all agreed that playing a series of syllables is better than playing single syllables one at a time.)
b. Move the matching label to the position of the stopping point.
c. Adjust the right margin of the "selection area" to the left and try to identify the stopping point of the second right-most syllable by listening to the audio.
d. Repeat the procedure "c" and "d" until the series of syllables are all correctly labeled.
e. Scroll the mouse wheel to move on to the next series of syllables. Then repeat "a", "b", "c", "d", "e" until the whole file is finished.

3.6.7 Difficult Cases for a Researcher

a. Hard to Follow
Sometimes a speaker has an accent or speaks too rapidly, which will make a researcher unable to distinguish the syllables from one another. (In the green rectangle of Figure 39)
b. "Flat" Waveform
The envelope of the waveform can often help a researcher estimate the position of syllables. However, in some files, speakers speak very quietly or the background noises are too loud, which makes the waveform envelope nearly flat. A researcher needs to listen much more times to locate the syllables. (In the red rectangle of Figure 39)
c. Serious Positional Problems
Sometimes a label was misplaced nearly 100 syllables away from where it should be.

3.6.8 Easy Cases for a Researcher

a. Mild Positional Problems
Some positional problems are not serious. Sometimes a misplaced label is only one drag away from the correct position.
b. Interesting Speech
When a female researcher listens to speech about makeup (or any other interesting topic) or when a male researcher listens to speech about videogames, it is less boring for the researcher to do aligning.

3.6.9 Headphones

The five student researchers used high quality headphones, with brands and model numbers as indicated in the table below.

Table 25: Summary of headphones used for listening to audio

Researcher	Brand	Model Number
Chen, Xiao	Sennheiser	PX80
Mao, Mao	Logitech	
Pan, Xu	Philips	SHE3590
Zhang, Hao	HP	BHP-95NCV
Zhang, Zifan	Sennheiser	HD 280 Pro

4 Results and Discussions

Each subsection of the main body of the report has its own results and discussion. Automatic classification and automatic recognition experiments were completed for Mandarin lexical tone identification. The highest classification accuracy obtained is about 76%. Highest recognition accuracy obtained was about 71%. These results compare favorably with the accuracy of native human listeners for tone recognition, when only limited context is available. In other work, with phone recognition for base syllables approximately 69% phone recognition accuracy was obtained. Complete character recognition has not yet been done.

5 Conclusion

Each subsection of the main body of the report has its own conclusion.

6 References:

[1] Wikipedia: “List of language by number of native speakers” http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] Wang, H.M., Shen, J.L., Yang, Y.J., Tseng, C.Y., and Lee, L.S., (1995). "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data," Proc. ICASSP 1995, Vol 1, pp.61-64.

[3] Zahorian, S.A, Silsbee, P., and Wang, X., (1997). “Phone classification with segmental features and a binary-pair-partitioned neural network classifier,” Proc. ICASSP 1997, pp.1011-1014.

[4] Zahorian, S. A., and Hu, H., (2008). “A spectral/temporal method for robust fundamental frequency tracking,” Journal of the Acoustic Society of America, Vol. 123, Issue 6, pp. 4559-4571.

[5] Huang, H. C.-H. and Seide, F. (2000). "Pitch tracking and tone features for Mandarin speech recognition." Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), Vol.3, pp.1523-1526.

[6] Liu, Z., Zhang, P., Shao, J., Zhao, Q., Yan, Y. and Feng, J. (2007). "Tone recognition in mandarin spontaneous speech." Proc. of the 4th International Conference on Non-Linear Speech Processing (NOLISP 2007).

[7] Chang, E., Zhou, J.L., Di, S., Huang, C., and Lee, K.F., (2000). "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," ICSLP '00.

[8] Zhou, J.L., Tian, Y., Shi, Y., Huang, C., and Chang, E., (2004). "Tone articulation modeling for Mandarin spontaneous speech recognition." Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004), Vol. 1, pp. 997-1000.

[9] Wang, C., and Seneff, S., (2000). "Improved tone recognition by normalizing for coarticulation and intonation effects," Proc. ICSLP 2000, pp.83-86.

[10] Lin, W.-Y., and Lee, L.-S., (2003) "Improved tone recognition for fluent Mandarin speech based on new inter-syllabic features and robust pitch extraction," IEEE 8th Automatic Speech Recognition and Understanding Workshop, PP.237-242.

[11] Chang, E., Zhou, J. L., Do, S., Huang, C., and Lee, K. F., “Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones,” 2000, ICSLP 2000.

[12] Zhou, J. L., Tian, Y., Shi, Y., Huang, C., and Chang, E., "Tone articulation modeling for Mandarin spontaneous speech recognition," Proc. ICASSP 2004, Vol. 1, pp. 997-1000.

[13] Wang, S., W., and Levow, G., "Improving Tone recognition with Combined Frequency and Amplitude Modeling," Interspeech, ICSLP, 2006.

[14] Li, A., and et al., "RASC863-A Chinese Speech Corpus with Four Regional Accents," Chinese Academy of Social Sciences technical report, 2004.

[15] "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium (LDC) report.

[16] S.A. Zahorian, Hongbing Hu, Zhengqing Chen, Jiang Wu, “Spectral and Temporal Modulation Features for Phonetic Recognition,” Interspeech 2009.

[17] O. Kalinli, “Tone and Pitch Accent Classification Using Auditory Attention Cues,” IEEE ICASSP, pp. 5208-5211, 2011.

[18] X. Lei, M. Siu, M.Y. Hwang, M. Ostendorf and T. Lee, “Improved Tone Modeling for Mandarin Broadcast News Speech Recognition,” IEEE ICASSP, pp. IV - 665 - IV - 668, 2007.

[19] Y. Lai and S.H. Wu, “The Effect of Segmental Makeup on Mandarin Word and Tone Recognition,” meeting of Acoustical Society of America, Kansas City, 2012.

[20] R. Wayland, D. Laphasradakul, E. Kaan, R. Cao, “Perception of Pitch Contours among Native Tone Listeners,” Interspeech 2012, Wed.08c.01, Portland, 2012.

[21] S. Promo-on, F. Liu, and Y. Xu, “Post-low Bouncing in Mandarin Chinese: Acoustic Analysis and Computational Modeling,” Journal of Acoustical Society of America , pp. 421-432, 2012.

[22] J. Zhou, T. Ye, Y. Shi, C. Huang and E. Chang, “Tone Articulation Modeling For Mandarin Spontaneous Speech Recognition,” IEEE ICASSP , 2004.

[23] S. B. D and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoustic., Speech, Signal Processing, vol. ASSP- 28, no. 4, pp. 357-366, 1980

[24] Md. Afzal Hossan, S. Memon, M A Gregory, “A novel approach of MFCC feature extraction,” IEEE Trans. On Signal Processing and Communication 2010 4th international conference.

[25] Wu Junqin, Yu Junjun, “An Improved Arithmetic of MFCC in Speech Recognition System,” IEEE 201, pp 719-722

[26] S.A. Zahorian, Silsbee, P., and Wang, X., “Phone Classification with Segmental Features and a Binary-Pair partitioned Neural Network Classifier,” Proc. ICASSP 1997, pp.1011-1014, 1997.

[27] M. Karjanadecha and S.A. Zahorian, “Signal Modeling for High-Performance Isolated Word Recognition,” IEEE Trans. on Speech and Audio Processing, 9(6), pp.647-654, 2001.

[28] S. Strahl, “Analysis and design of Gammatone signal models,” J. Acoust. Soc. Am. 126, pp. 2379-2389, 2009.

[29] B. Moore, R. Peters, and B. Glasberg, “Auditory filter shapes at low center frequencies,” J. Acoust. Soc. Am. 88, 132–140, 1990.

[30] B. Moore and B. Glasberg, “A revision of Zwicker’s loudness model,” Acta. Acust. Acust. 82, 335–345, 1996

[31] Holdsworth J. et al. “Implementing a Gamma Tone Filter Bank,” in SVOS Final Report – Part A: The Auditory Filter bank, MRC Applied Psychology Unit, Cambridge, England, 1988.

[32] L. Rabiner, B.H. Juang, “Fundamentals of speech Recognition,” Prentice Hall Signal Processing Series, 1993.

[33] Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, “Speech Recognition Using MFCC,” International conference on computer graphics, simulation and modeling (ICGSM’2012) july 28-29, 2012 Pattaya (Thailand).

[34] Ms. Sahidullah, Goutam Saha, “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition,” Speech Communication 54(4): 543-565 (2012)

List of symbols, abbreviations and acronyms

AFRL	Air Force Research Laboratory
ASR	Automatic Speech Recognition
DCT	Discrete Cosine Transform
DCTC	Discrete Cosine Transform Coefficient
DCS	Discrete Cosine Series
DCSC	Discrete Cosine Series Coefficient
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
LDC	Linguistic Data Consortium
MFCC	Mel-Frequency Cepstral Coefficient
NN	Neural Network
RASC863	Regional Accented Speech Corpus by National 863 Project